Spring 5-19-2018

Computational Theories For Human Stereo Vision

Han Gao
gaoh@smu.edu

COMPUTATIONAL THEORIES FOR HUMAN STEREO VISION

Approved by:

Dr. Carlos Davila
Associate Professor of Electrical

Engineering

Dr. Scott Douglas
Professor of Electrical Engineering

Dr. Dario Villarreal
Assistant Professor of Electrical

Engineering

Dr. Prasanna V Rangarajan
Research Assistant Professor of Electrical

Engineering

COMPUTATIONAL THEORIES FOR HUMAN STEREO VISION

A Thesis Presented to the Graduate Faculty of the

Lyle School of Engineering

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Master of Science in Electrical Engineering

with a

Major in Electrical Engineering

by

Han Gao

B.A., Electrical Engineering, South China University of Technology, Guangzhou

May 19, 2018

ACKNOWLEDGMENTS

This work could not have been accomplished without the wisdom of my advisor, Prof. Carlos Davila, along with his many colleagues in the Department of Electrical Engineering here at SMU. I'm also forever grateful to my family for being so patient with me and my friends for keeping encouraging me when I feel down.

Gao, Han B.A., Electrical Engineering, South China University of Technology, Guangzhou

Computational Theories for Human Stereo Vision

Advisor: Dr. Carlos Davila

Master of Science in Electrical Engineering degree conferred May 19, 2018
Thesis completed May 01, 2018

Binocular stereopsis refers to the ability to perceive depth, which has always been a central problem in perception since the time of da Vinci. The foremost theoretical difficulty that arises when attempting to understand how the visual system computes disparity is known as the correspondence or matching problem. Decades of research upon macaque primary visual cortex has shown that in each layer of the primary visual cortex (V1) long-range horizontal connections among striate cortex cells exist which integrate information from different parts of the visual field [1]. Inspired by long-range horizontal connections in V1 and the Jeffress model [2], a time-delay neural network which represents a time difference spatially to solve the sound localization problem, we propose a dynamic computational stereo matching algorithm that predicts how the visual system solves the stereo matching problem using left-eye and right-eye images. In our model, eye movements like saccades and drift, transform spatial information into time domain signals. A neural structure similar to the Jeffress model is used to decode disparity. To enhance performance, we introduce Gabor filters whose two-dimensional functions have been proven to be a good fit to the receptive field (RF) profiles of simple cells in the striate cortex [3,4]. Further, we fitted our model with the Combination of Receptive Fields (CORF) model which is a computational model for the lateral geniculate nucleus (LGN) cell with center-surround receptive fields (RFs) proposed by Azzopardi, G., and Petkov, N. [5]. Both random-dot stereograms (RDS) and natural stereo images were used for testing. The results indicate that our model is a possible solution for the stereo matching problem, but more details need to be added for better performance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Page

To you as a reader.

Chapter 1

Introduction

Humans can obtain an unambiguous perception of depth with one eye or when viewing a pictorial image of a 3-dimensional scene [15]. Nevertheless, when using both eyes to view a real scene, the perception becomes most qualitatively vivid. Scientists refer to this perception phenomenon as stereopsis.

Stereopsis has been an area of interest among the perception research community since the time of the Italian Renaissance. Aiming to improve the perception of depth in pictorial images, many Renaissance artists including da Vinci began to explore human vision. A major puzzle in their research was why it was impossible to reproduce the vivid impression of depth obtained in real scenes, despite accurate and realistic perspective rendering [16].

In 1838, Wheatstone identified a connection between this qualitative impression and the feature differences between monocular vision while viewing real objects (binocular disparities). Based on his observation, Wheatstone invented the stereoscope, which can produce binocular disparities by presenting two pictures to each eye with slight differences [17]. The excellent vividness of depth sensed under stereoscopic viewing led to the widely accepted opinion that binocular disparity was the causal basis of the perception of depth, and the term stereopsis has come to simply imply depth perception from binocular disparity [15]. After centuries of research, all kinds of hypotheses aiming to explain the mechanism embedded in the human vision system can be divided into three main types [15]:

1. Binocular disparity hypothesis: The most widely accepted hypothesis, it states binocular vision is the main reason for stereopsis. Depth perception happens when binocular disparities can represent the perceived depth relations.

2. Cue coherence/depth magnitude hypothesis: This hypothesis states that stereopsis is related to the magnitude of perceived depth. When cues coherently specify similar

depth values, a greater and more accurate magnitude of depth is perceived, resulting in a strong impression of stereopsis. The conflict between depth cues results in a diminished magnitude of perceived depth and the lack of stereopsis [15].

3. Visual parallax hypothesis: In this hypothesis, scientists indicate stereopsis is the outcome of visual parallax. Parallax can be obtained from the differences between left eye vision and right vision, or between different images in binocular vision.

In this thesis, we will focus on the first hypothesis, decoding depth from binocular disparities.

Figure 1.1. A wooden stereoscope with a pair of natural stereo images.

In the rest part of this chapter, we are going to give a brief introduction to the background knowledge. We will present several famous computational stereo matching models in the second chapter. In chapter three, the dynamic stereo matching algorithms we proposed will be described in detail. Simulation results and discussions will be given in the latter part.

1.1. Background

1.1.1. Binocular Disparity

Our brains usually receive two similar images taken from two nearby points at the same horizontal level because of the way our eyes are positioned and controlled [6]. Marr and Poggio [6] concluded in their paper that binocular disparity measurement consisted of three steps: Step 1: from one image, select a specific point on a surface in the scene. Step 2: identify the same point from another image. Step 3: measure disparity between the two corresponding points. So the main problem for the disparity calculation comes in finding corresponding points from monocular images. However, in practice as illustrated in Figure 1.2, it is hard to do it precisely because of false matches. [6].

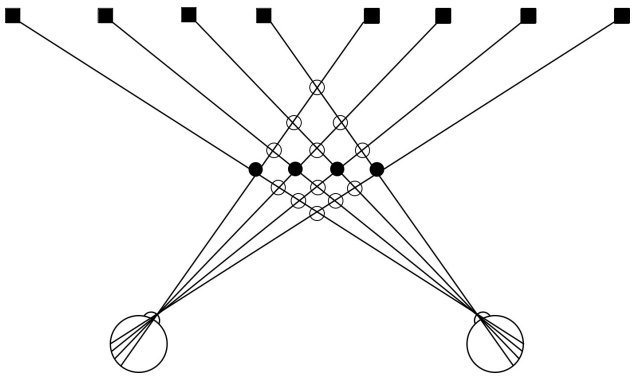

Figure 1.2. Binocular correspondence ambiguity. When viewing multiple targets (black squares with zero disparity), there are many possible binocular matches, in this case, 16 matches in total (intersections between each line). Among these possible matches, only four are correct matches (•), and the other 12 are false matches (∘). Redrawn from Marr and Poggio [6].

In order to formulate the correspondence calculation accurately, two local constraints were proposed based on physical rules [11]:

1. Uniqueness. At most one disparity value can be given to each point from each image. This is because a particular point from a physical surface at one time can only have one position in space.

2. Continuity. Disparity varies smoothly along the surface most of the time. This constraint is a consequence of the fact that surfaces are generally smooth. Also, only the boundaries in an area with small fraction are discontinuous in depth.

These rules are widely used by many stereo matching algorithms. We also apply these constraints to our model for false matches elimination.

1.1.2. Eye Movements

When we are reading or viewing a scene or searching for an object in the surrounding environment, we are able to do this by moving our eyes across the visual field. The fact that eye movements occur so frequently in various visual cognition tasks (like reading and scene perception) has led a number of research investigators to utilize the eye movement record to infer something about the cognitive processes involved in such tasks (Figure 1.3) [18].

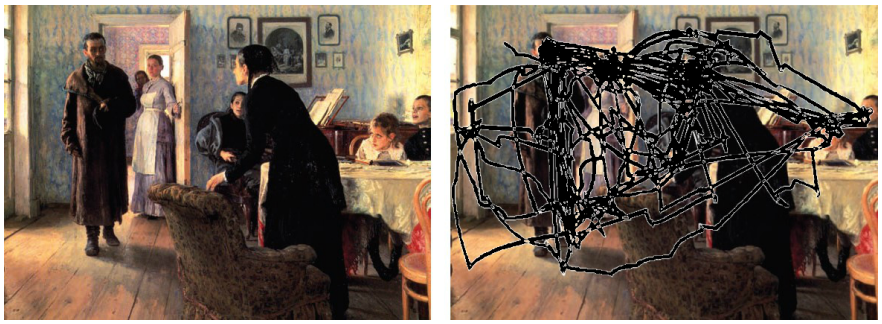

Figure 1.3. Eye movements during vision. Left image shows the original painting (An Unexpected Visitor by Ilya Repin) which was presented to a human subject for several minutes. Right image shows the saccades (black lines) made by the subject to selected fixation points (spots). Figure reproduced from [7].

In general, eye movements can be divided into four types including saccades, smooth pursuit movements, vergence movements, and vestibulo-ocular movements. Saccades are rapid, ballistic movements of the eyes that abruptly change the point of fixation which are used to change our view of the world [8] (ballistic movements refer to the high-velocity, short-time-period muscle contractions). The amplitude range for saccades is relatively large which varies from small movements made when reading, for example, to much larger movements made when looking around a room [8]. We can elicit saccades voluntarily, but most of the time, saccades happen unconsciously whenever our eyes are open, even when we intently stare at a target of interest. Saccades are also the rapid eye movements which happen during

an important stage of sleep. Figure 1.4 shows the time course of a saccade eye movement. In this example, it takes 200 ms for saccade eye movements to start after the targets abruptly moves to the right. Research indicates that saccade eye movements are ballistic because the saccade-generating system cannot respond to the subsequent changes in the position of the target during the course of the eye movement [8]. If the position of the target changes again during this time (on the order of 15-1100 ms), the saccades will miss the target, and another saccade will be elicited to correct this error [8].

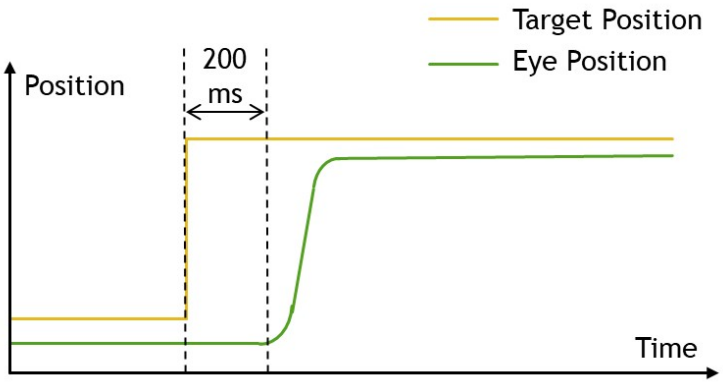

Figure 1.4. The metrics of a saccade eye movement. The position of the fovea is given by the red line. The position of a fixation target is given by the blue line. When we move the target abruptly to the right, it takes the eye 200 ms to make a saccade. Redrawn from [8] (Figure 20.4).

.

Smooth pursuit movements are much slower tracking movements of the eyes designed to keep a moving stimulus on the fovea [8]. As shown in Figure 1.5, smooth pursuit movements are controlled voluntarily, which means that the observer can decide whether or not to track a moving stimulus. However, for most people making a smooth pursuit movement without a moving target is hard which usually ends up with a saccade, and only highly trained people can do it.

Vergence movements align the fovea of each eye with targets located at different distances from the observer [8]. Almost all the eye movements are conjugate eye movements (two eyes moving in the same direction), except for vergence movements which are disconjugate (or

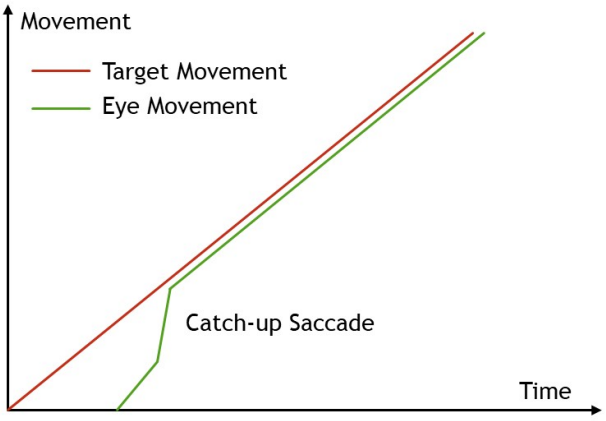

Figure 1.5. The metrics of smooth pursuit eye movements. After a quick saccade to capture the target, the eye movements (green line) reach the velocity of the moving target (red line) for stimulus tracking. Figure from [8] (Figure 20.5).

.

disjunctive) eye movements. They are responsible for either convergence or divergence of the lines of sight of each eye to capture an object which is nearer or farther away [8].

Vestibulo-ocular movements enable our eyes to be stable relative to the external world by compensating for head movements. In other words, when the position of head varies, vestibulo-ocular movements keep the visual image from "sliding" on the retina surface. We can appreciate the action of vestibulo-ocular movements by moving our heads from side to side while fixating on a target. The eyes automatically compensate for the head movement by moving the same distance but in the opposite direction, thus keeping the image of the object at more or less the same place on the retina [8].

1.1.3. Role of Eye Movements in Stereopsis

The question of the role of eye movements during stereoscopic vision has been bothering biologists for a long time, and numerous hypotheses have been proposed. Brucke held the opinion that when we were viewing an object, a rapid series of eye movements were made during the course of which the various parts of the target are successively fixated [19].

In 1841, Dove performed experiments which he believed completely disproved the eye

movement theory. In Dove's experiment, he was able to exclude the possibility of eye movements by using a brief exposure time of the stimulus [19]. His experiments have been repeated by many other researchers and numerous retinal pattern theories of stereopsis were proposed [19–21]. In 1958, Ogle and Weil when they repeated Dove's experiments put forward a concept called stereoscopic threshold which referred to the disparity difference yielding 75 percent or 83 percent correct responses. They summarized the results as follows [21]:

1. With constant luminance adaptation, the contrast of the test object has relatively little effect on the stereoscopic threshold.In other words, stereoscopic acuity is independent of the luminance for a given level of luminance of the adapting background.

2. With constant luminance adaptation, the stereoscopic threshold increases four times when the exposure duration of the test object decreases 1.08 seconds. Therefore, stereoscopic acuity rapidly exacerbates when the exposure duration of the test object decreases. The threshold tends to be related to duration length as an exponential function of the exposure time [21]. When the exposure time goes down, the threshold increases slowly at first and then becomes more and more quickly.

3. The increase in the threshold with decrease in exposure time appears to be independent of "whether or not both reference and test objects are seen with disparate images" [21].

Based on their conclusions, Ogle *et al* made the assumption that the stereoscopic acuity was "enhanced in some relationship to the number and extent of the excursions of these involuntary eye movements" [21] which took place during the exposure of the object. For very short exposure which is shorter than $\frac{1}{200}$ seconds, the retinal image is "stopped". Correspondingly they believed that there would be an upper threshold of stereopsis. Their assumption was entirely consistent with the results that the threshold was independent of the contrast and the disparities of the stereo images [21].

Washburn illustrated in his paper that the vivid perception of depth is a distinctly motor experience. He also suggested the possibility that binocular rivalry played an important role in normal binocular vision [19]. Binocular rivalry is a type of multistable visual phenomenon

in which perception alters between different stimuli presented to corresponding regions of the left and right eye.

One of the most famous experiments was proposed by Clark, and the results of his examinations led to the conclusion that stereopsis could be explained fundamentally as the result of the cerebral organization and a series of unified intermittent retinal impressions, which happened as clear vision changed from one eye to another [19]. Eye movements during fixation were the causes of these oscillations.

Although significant effort has been made to determine the role of eye movements in stereopsis, the mechanisms behind this are still unclear. However, we believe that eye movements are the indispensable part of stereopsis and consider eye movements as the theoretical foundation of our computational models.

1.1.4. Interaural Time Differences and The Jeffress Model

Many animals use interaural time differences (ITDs) to locate the source of low-frequency sounds (we are capable of using ITDs as small as 10 to 20 microseconds to distinguish directional differences of sound sources' locations) [22]. In 1948, Jeffress proposed a computational model which can represent a time difference spatially to account for the neural mechanisms of ITD detection [22]. This model is known as the Jeffress model or the place theory. The Jeffress model (Figure 1.6) was proposed based on three fundamental assumptions [22]:

1. The ascending nerve fibers' conduction times are arranged in an orderly fashion, and serve as delay lines.

2. Input signals synchrony is converted into output firing rate by coincidence detectors.

3. In the cell array, a neuronal place map is formed based on the systematic variation in firing rate.

In general, the Jeffress model is a time-delay neural network which uses a set of delay lines and coincidence detectors to compute a temporal cross-correlation between two sets of input signals from left and right pathways. While decades have passed, the Jeffress model has

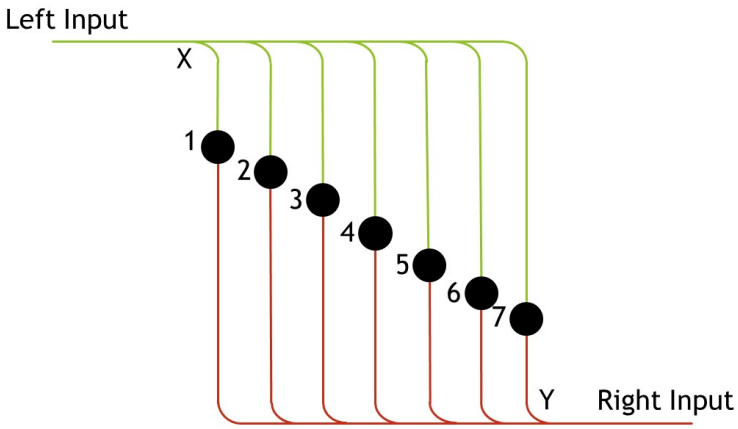

Figure 1.6. The Jeffress model with delay lines (red and green lines) and coincidence detectors (black dots 1-7). Point X and Y receive the signals at the same time. In this case, coincidence detectors (1-7) are tuned to seven different disparities from -3 to +3. For example, if the disparity between two input signals is +1, then only coincidence detector 5 can receive two inputs simultaneously and has the strongest response. Redrawn from [2].

become a dominant model for solving the ITD problem and researchers have found similar neuronal ITD maps in many different kinds of animals [22]. However, recent research upon mammals indicates that the Jeffress model is not the only way for sound localization [22].

1.1.5. Receptive Fields of Retinal Ganglion Cells

C. Sherrington created the term *receptive field* (RF) in 1906 when he was analyzing the scratch withdrawal reflex: "The whole collection of points of skin surface from which the scratch-reflex can be elicited is termed the receptive field of that reflex" [9]. Years later when it was possible to record signals from single neurons in the eye, H. Keffer Hartline used the concept of the receptive field during his research upon the horseshoe crab's retina, "The region of the retina which must be illuminated in order to obtain a response in any given fiber . . . is termed the receptive field of that fiber" [9].

Each small window on visual space represents a retinal ganglion cell's receptive field in the visual system. The receptive field of a retinal ganglion cell consists of different numbers

of photoreceptors for different locations on the retina. A receptive field which is close to the fovea contains fewer photoreceptors and covers a smaller area, while a receptive field which is farther away from the fovea contains more photoreceptors and covers a larger area. To reach the photoreceptors which are located at the back of the retina, light has to pass through nerve cell layers. Then, signals from the photoreceptors are transmitted to a retinal ganglion cell through "neurons in the outer and inner nuclear layers" [9].

However, their research upon the cell's receptive field was limited, as only one spot of light was used. Both Hartline and S. Kuffler used two small spots of light to study the mammalian retina and found a lateral inhibitory region in the receptive field. In 1953, Kuffler observed that "not only the areas from which responses can actually be set up by retinal illumination may be included in a definition of the receptive field but also all areas which show a functional connection, by an inhibitory or excitatory effect on a ganglion cell" [9]. Therefore, Kuffler concluded that the receptive fields of ganglion cells had a center-surround organization with functionally distinct subareas, which could be divided into two categories: on-center and off-center (as shown in Figure 1.7). And following research showed that neurons in LGN had similar receptive fields [9].

On-center cells fire when a spot of light is provided within a circular central region. To the contrary, off-center cells fire when a spot of light in the center is turned off (Figure 1.8). The surrounding annular region has the opposite sign. For on-center cells, a light stimulus on the surrounding area will produce a response when the light is turned off, a response termed on-center, off-surround [9]. The center and surround areas are mutually inhibitory. If a diffuse light is given to both center and surround areas, there will be little or no response. In turn, a light-dark boundary which moves across the center-surround receptive field can produce a brisk response. As these neurons are more sensitive to boundaries and contours (differences in illumination) than to uniform surface, they are believed to encode information about the contrast in the visual field [9].

1.1.6. The Structure of Primary Visual Cortex

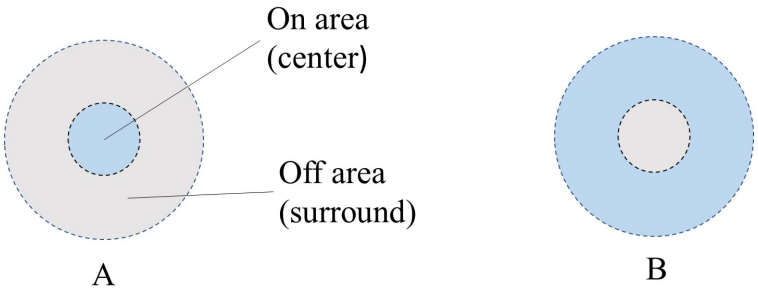

Figure 1.7. Receptive fields of retinal ganglion cells with center-surround structure. A. On-center structure. B. Off-center structure.

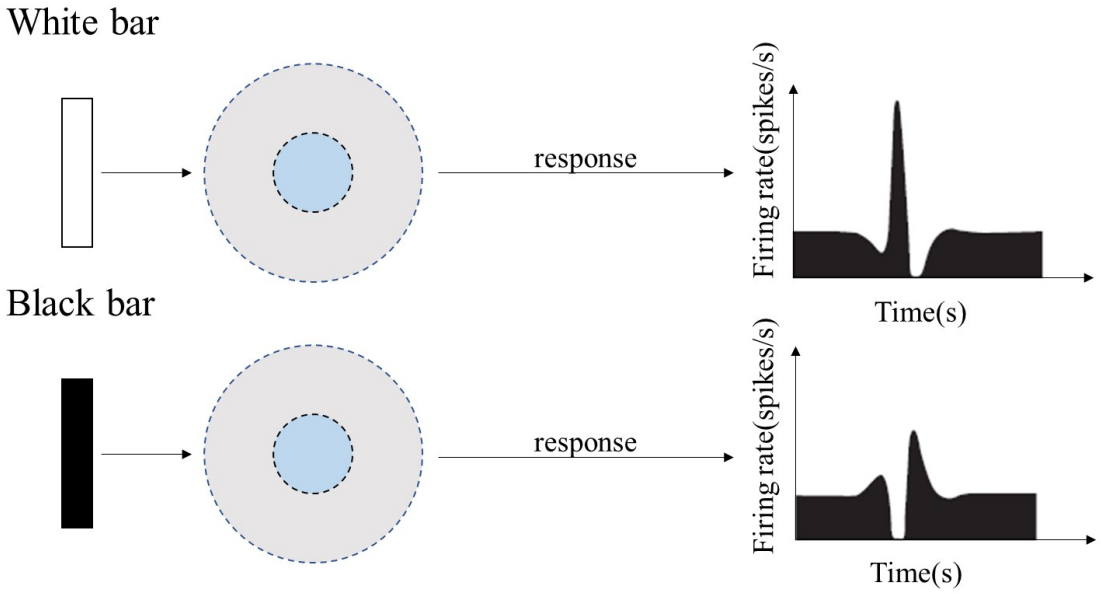

Figure 1.8. The representation of moving objects by retinal ganglion cells. On-area is sensitive to high luminance stimulus and will be inhibited by low luminance stimulus, off-area has exactly reverse response pattern of on-area. Figure reproduced from [9] (Figure 25-8).

The primary visual cortex, also known as V1, is a part of the cerebral cortex which is located in the occipital lobe. "The dominant feature of the functional organization of the primary visual cortex is the visuotopic organization of its cells: the visual field is systematically represented across the surface of the cortex" [9]. Notice that in the primary visual cortex, cells with similar functional properties are closely located together in columns which stretch from the cortical surface to the white matter. The columnar structures are related to the functional properties which were analyzed in any given cortical area and therefore reflect the functional role of that area in vision [9]. These properties developed in V1, including orientation selectivity and the integration of binocular inputs, is measured as the relative strength of input from each ocular dominance (eye). Ocular-dominance columns (Figure 1.9) indicate that thalamocortical (cortex of thalamus) inputs from different layers of the lateral geniculate nucleus (LGN) which is located in thalamus are segregated with each other [9].

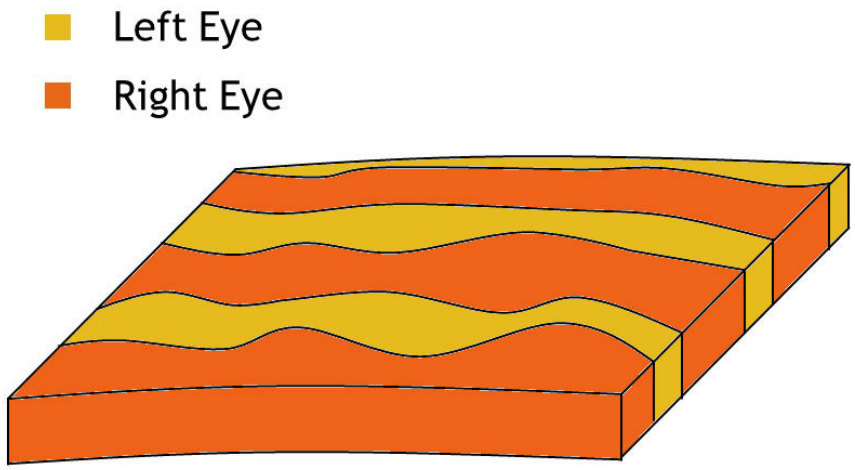

Figure 1.9. The ocular dominance columns in the primary visual cortex. The yellow and orange stripes represent the surface view of the left and right ocular dominance columns. Figure reproduced from [9] (Figure 25-11).

Physiologists have already proved that in each hemisphere, the LGN receives input signals from the temporal retina of the ipsilateral eye and the nasal retina of the contralateral eye [9]. The LGN is a laminated structure which consists of two magnocellular layers (layers 1 and

2) and four parvocellular layers (layers 3 to 6) [9]. The inputs from the two eyes terminate in different layers:

- The contralateral eye sends input to layers 1, 4 and 6.
- The input signals from ipsilateral eye are sent to layers 2, 3 and 5.

The magnocellular and parvocellular layers are connected to separate sublayers of the primary visual cortex. The magnocellular layers project to sublayer IVCα and the parvocellular layers to sublayer IVCβ. Besides, the inputs from the contralateral and ipsilateral layers of the LGN are segregated into alternating ocular-dominance columns [9]. This segregation is maintained in the inputs from the LGN to V1, which produces the alternating right-eye and left-eye ocular dominance bands receiving inputs from the different layers of the LGN [9].

Neuron cells with similar orientation selectivity also combine into columnar structures as shown in Figure 1.10. Across the cortical surface, there is a "regular clockwise and counterclockwise cycling of orientation preference with the full 180^o cycle repeating every 750 μm" [9] Just like the ocular dominance columns, the orientation selective columns' periodicity varies from 750 to 1000 μm. And one full cycle of orientation columns is called a hypercolumn. The ocular dominance columns and orientation dominance columns are crisscrossed over the cortical surface.

Both types of columns were first mapped by recording the responses of neurons at closely spaced electrode penetrations in the cortex [9]. There are clusters of neurons located within the orientation and ocular dominance columns which are not sensitive to orientation but highly color selective. These units of color-selective neurons are distributed in a regular patchy pattern of blobs and interblobs which are a few hundred micrometers in diameter and 750 μm apart [9]. Blobs consist of clusters of color-selective neurons, which makes blobs specialized to respond to surfaces rather than edges.

With the new technique for axon labeling, a recombinant adenovirus bearing the gene for green fluorescent protein (GFP), Stettler, Das, and Bennett were able to map the long-range horizontal connections which were parallel to the surface of the primary visual cortex [23].

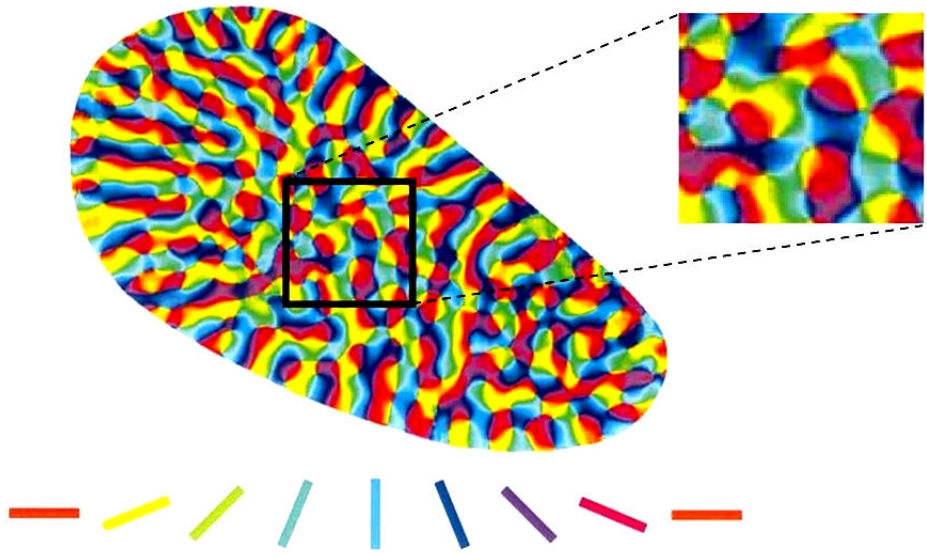

Figure 1.10. The structure of the orientation selective columns. Orientation selectivity at each location is color-coded according to the color code below. Figure adapted from [10].

These connections and their functions in the primary visual cortex were first found and analyzed by Gilbert and Wiesel, who used intercellular recordings and dye injection to correlate anatomical features with cortical function [9]. Stettler *et al* proved that the orientation selectivity of the intrinsic connections is consistent with orientation dependence of correlating firing, flank facilitation, and contour saliency [23]. More importantly, the orientation selective columns which were separated by several millimeters have the strongest correlation with the columns with similar orientation preference [24]. The length of long-range horizontal connections, at most "7 mm end to end in cortical terms and 4^o visuotopically" [1], are well consistent with the scale of contextual interactions found in psychophysical and physiological investigations of contour integration [23].

These horizontal connections enable target neurons to integrate information over large parts of the visual field, as the organization of the visual cortex is visuotopically. Therefore, the long-range horizontal connections play an important role assembling the components of

a visual image into a unified perception.

The long-range horizontal connections between orientation-selective columns beg the question that does the primary visual cortex have similar structures to the Jeffress model which are responsible for binocular disparity encoding? This question is the first motivation for this thesis.

Chapter 2

Computational Models for Stereo Matching

2.1. A Cooperative Algorithm For Stereo Matching

In 1976, D. Marr and T. Poggio proposed their famous cooperative algorithm to extract disparity information from stereo image pairs [11]. As we mentioned before, Marr and Poggio illustrated in their paper that three steps were involved in perceiving stereo disparity: Step 1: from one image, select a specific point on a surface in the scene. Step 2: identify the same point from another image. Step 3: measure disparity between the two corresponding points [11]. Particularly, two constraints were introduced which could be translated into two rules for how monocular descriptions are combined. The first rule is Uniqueness, it stipulates each item from each image may be assigned at most one disparity value. The second rule is Continuity, it assumes in most cases on the surface of one scene disparity changes smoothly. Their cooperative algorithm for the computation was derived by constructing an explicit representation of the two rules [11]. The structure of a network for implementing the cooperative algorithm in the one-dimensional case is shown in Figure 2.1 [11]. If we put cells (one-dimensional case) shown in Figure 2.2a at each intersection and connect them in the way that inhibitory connections are along the thick vertical and horizontal lines and excitatory connections are along the diagonal dash lines, then, given the parameters are appropriate, "the stable states of such a network will be precisely those in which the two rules are obeyed" [11]. By expanding the local excitatory neighborhood to two-dimensional, we can fit this idea with two-dimensional images (Figure 2.2b). A simple form of the cooperative algorithm can be expressed by the following set of difference equations [11]:

$$C_{xyd}^{(n+1)} = \sigma\{\sum_{x`y`d`\in S(xyd)} C_{xyd}^{(n)} - \epsilon \sum_{x`y`d`\in O(xyd)} C_{xyd}^{(n)} + C_{xyd}^{(0)}\} \tag{2.1}$$

where $C_{xyd}^{(n)}$ is the state of the cell at position (x, y) with disparity d at iteration n. n is time step, S and O are the circular (excitatory) and thick line (inhibitory) neighborhoods of the cell xyd in Figure 2.2b, ϵ is the inhibition constant, and σ represents a sigmoid function with range [0,1] [11].

As one of the most famous computational algorithms for stereo matching, the main advantage is their algorithm does not require a specific minimum or maximum correlation area (characteristic scale) which would restrict analysis. The connection between the early stage vision perception in human brain and their cooperative algorithm was unknown even for authors.

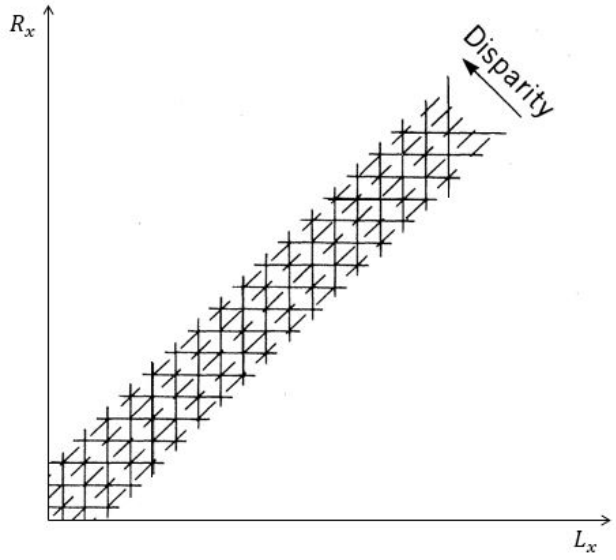

Figure 2.1. The one-dimensional structure of the two rules. L_x and R_x represent the location of features on the left and right eye images. The lines of sight from left eye and right eye are represented by the vertical and horizontal thick lines [11]. The intersections of them are possible matches which correspond to different disparity values. The points along the diagonal dash lines have the same disparity value. Redrawn from [11] (Fig.2).

2.1.1. Global Consistency Constrain Model

Based on D. Marr and T. Poggio's work, N. Sato and M. Yano put forward a model of stereopsis with an explicit global measure of correspondence [25]. In their experiments, N.

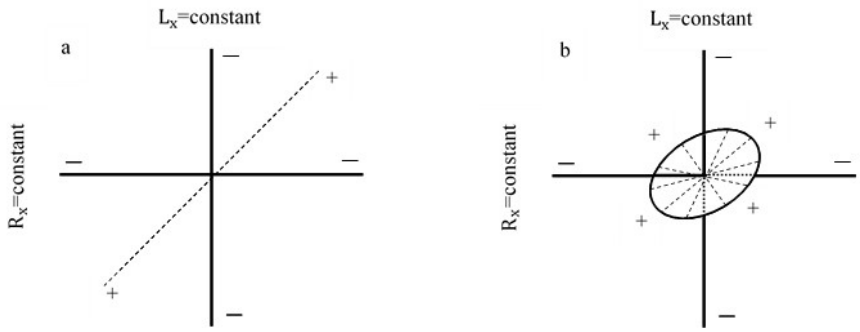

Figure 2.2. The local structure at each cell of the network for implementing the cooperative algorithm which is described by Eq 2.1. Dash lines represent the excitatory connections and solid lines represent inhibitory connections. The one-dimensional case is shown in a and two-dimensional case is shown in b. Redrawn from [11] (Fig.2).

Sato and M. Yano found although cooperative algorithm implements the uniqueness constraint and the continuity constraint to solve binocular correspondence problem, they are not sufficient to define the proper correspondence uniquely [25]. So, they set cooperative algorithms continuity constraint as a local constraint and added a global estimation of correspondence to enhance its robustness in Random-dot stereograms (RDSs) matching problem. Their model is a hierarchical system with two layers [25]:

- The first layer is the monocular layer which contains two monocular layers representing left eye vision and right eye vision, and each monocular layer consists of a 3232 array of complex cells [25].
- The second layer is the integration layer which consists of seven disparity planes representing disparity -3,, +3 [25]. Specifically, their model provides a global measure of the correspondence at each position [25].

The activity x of a single cell at (i, j) is given by following equation [25]:

$$\frac{d^2x_{ij}^L}{dt^2} + (A_1 y_{ij}^{L^2} + B_1 x_{ij}^L + C_1)\frac{d^2x_{ij}^L}{dt} + (A_2 y_{ij}^{L^2} + B_2 x_{ij}^L + C_2)x_{ij}^L = d_{ij}^L \tag{2.2}$$

$$\frac{d^2x_{ij}^R}{dt^2} + (A_1 y_{ij}^{R^2} + B_1 x_{ij}^R + C_1)\frac{d^2x_{ij}^R}{dt} + (A_2 y_{ij}^{R^2} + B_2 x_{ij}^R + C_2)x_{ij}^R = d_{ij}^R \tag{2.3}$$

where L and R denote the left and right monocular layers, parameter A, B, C are constants and d_{ij}^L, d_{ij}^R represent the input for cells at position (i, j) in left and right monocular layers, which are given by [25]:

$$d_{ij}^L = I_{ij}^L\{D + \sum_{l=-6, l\neq 0}^{6} \alpha_{ij}^L(l) + \sum_{k=-3}^{3} \beta_{ij}^L(k) + \sum_{k=-3}^{3} C_\gamma \gamma_{ij}^L(k)\beta_{ij}^L(k)\} \tag{2.4}$$

$$d_{ij}^R = I_{ij}^R\{D + \sum_{l=-6, l\neq 0}^{6} \alpha_{ij}^R(l) + \sum_{k=-3}^{3} \beta_{ij}^R(k) + \sum_{k=-3}^{3} C_\gamma \gamma_{ij}^R(k)\beta_{ij}^R(k)\} \tag{2.5}$$

where I_{ij}^L and I_{ij}^R are the value of dotes at position (i, j) in the left and right RDSs, which are inputs of the cell at (i, j) in the left and right RFs. Function $\alpha_{ij}(l)$ is the correspondence between the cell at (i, j) and the cell at $(i + l, j)$ in one monocular layer, $\beta_{ij}(k)$ denotes the interaction between the cell at (i, j) and the cell in the opposite layer of disparity k. More importantly, $\gamma_{ij}(k)$ is the global measure of correspondence at (i, j) of disparity k, and it is given by [25]:

$$\gamma_{ij}(k) = \frac{s_{ij}^k}{\sum_{l=-3}^{3} s_{ij}^l} \tag{2.6}$$

where s_ij is the activity at (i, j) in the integration layer of disparity k. From Equation 2.6, we can see when $\sum_{l=-3}^{3} s_{ij}^l$ is large, small s_{ij}^k (k=-3,,3) can be neglected and $\gamma_{ij}(k)$ only depends on large s_{ij}^k which is because the pimping to the activity at a position near (i, j). When the quantity $\sum_{l=-3}^{3} s_{ij}^l$ is small, then $_{ij}(k)$ depends on both small s_{ij}^k and big s_{ij}^k which is caused by the pimping to the activity at a position distant from (i, j). This indicates that the pumping to the activity at a position near (i, j) can be equivalent to one at a position distant from (i, j), depending on the conditions of the integration layers [25].

The authors concluded that as the global measurement was independent of the size of binocular objects, and the global measurement could reasonably represent the degree of correspondence between two monocular layers [25]. With the help of the global measurement of correspondence, this model can handle both the superimposed surface and the ambiguously perceivable surface which is a big improvement compared with old models.

2.2. Energy Model

2.2.1. Simple Cells

In 1959, the simple cells which were observed in the cat's striate cortex were firstly used to describe binocular interactions by Hubel and Wiesel [26]. In their experiments, they found that stimulating on or off subregions of the left and right eye receptive fields at the same time resulted in response summation, whereas stimulating an on subregion in one eye and off subregion in another eye canceled the response, which indicated that the binocular interaction of signals might be linear. They also found that some cells only respond to binocular stimulus, which suggested that there was a nonlinear binocular interaction. Cells selective to binocular disparity were also found in the macaque monkey striate cortex by many biologists and some of these cells were reported to respond to dynamic random-dot stereograms (RDS) (sensitive to binocular correlation) [3]. And cells with similar function (orientation selectivity) were also found in the macaque monkey primary visual cortex [27].

Based on the former studies, A. Anzai, I. Ohzawa, and R. D. Freeman were able to identify the underlying neural mechanisms for binocular interactions by determining the system structure of binocular simple cells and depicting the nature of nonlinearities in the system [3]. In their experiments, white noise was applied to analyze binocular interaction RFs and monocular RFs for simple cells in the cat's striate cortex. They also illustrated that most of the simple cells' binocular receptive fields were found to be proportional to the product of the left and right eye RFs [3]. Therefore the binocular disparity selectivity of simple cells is the result of their tuning for monocular phase in each eye and for binocular disparity [3]. The binocular interaction RF consists of a linear binocular filter and the following static nonlinearity which can be well modeled by a half-power function with an average exponent, like a half-squaring function [3], which was believed to play an important role in computations performed by simple cells.

A computational model was proposed to illustrate the functional roles of binocular simple cells. As given in Equation 2.7 (given $(W_L + W_R) > 0$, otherwise =0), simple cells sum the

outputs from the left and right eye linear filter and then the summations are rectified and squared [3].

$$(W_L + W_R)^2 = W_L^2 + W_R^2 + 2W_L W_R \tag{2.7}$$

As the output of the left and right eye linear filter is a weighted sum over space, Equation 2.7 can be rewritten as [3]:

$$\int L(X_L)S_L(X_L)dX_L + \int R(X_R)S_R(X_R)dX_R^{\ 2} = \int L(X_L)S_L(X_L)dX_L^{\ 2} + \int R(X_R)S_R(X_R)dX_R^{\ 2}$$

$$+ 2\int L(X_L)S_L(X_L)dX_L \int R(X_R)S_R(X_R)dX_R \tag{2.8}$$

where L and R represent the left and right eye RFs. Variables S and X are stimulus and position of stimulus. Based on the relationship between binocular interaction RFs and the product of monocular RFs, the third term in Equation 2.8 can be transformed into [3]:

$$2\int\int L(X_L)R(X_R)S_L(X_L)S_R(X_R)dX_L dX_R = \alpha \int\int B(X_L, X_R)S_L(X_L)S_R(X_R)dX_L dX_R \tag{2.9}$$

where B represents a binocular interaction RF and α is a constant. And by changing X_L and X_R to X and $X+\theta$, where θ is the corresponding binocular disparity, Equation 2.9 becomes [3]:

$$\alpha \int\int B(X, X+\theta)S_L(X)S_R(X+\theta)dX d\theta \tag{2.10}$$

Comparing Equation 2.10 with an interocular cross-correlation $\Phi_L R(\theta)$ of the stimulus [3]:

$$\Phi_L R(\theta) = \int S_L(X)S_R(X+\theta)dX \tag{2.11}$$

Obviously, Equation 2.10 is a weighted interocular cross-correlation integrated over all binocular disparities. Therefore, simple cells can perform interocular cross-correlation because of the static nonlinearity [3].

Quantitative physiological research has shown that two-dimensional Gabor functions which consist of complex sinusoidal carriers and Gaussian envelopes can well describe the binocular spatial RFs of simple cells [28]. For example, a vertically oriented Gabor function centered at origin can be given by the following equation [28]:

$$g(x, y, \phi) = \frac{1}{2\pi\sigma_x\sigma_y} exp(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}) cos(\omega x + \phi) \tag{2.12}$$

where ω is the preferred spatial frequency, σ_x and σ_y are the factors which determine the RF dimensions in x and y directions. ϕ represents the phase parameter for the sinusoidal carrier. Notice that, by rotating and translating this function respectively we can obtain other simple cells binocular RFs with different preference orientations and RF centers.

2.2.2. Complex Cells

Three years after their discovery of simple cells in the cat's striate cortex, Hubel and Wiesel demonstrated another type of cell in the cat's striate cortex, and they named it complex cell [26]. The original description of complex cells given by Hubel and Wiesel indicated that complex cells were nonlinear computing devices. In their experiments, they found that unlike simple cells, RFs of complex cells didn't have discrete on and off subregions. In fact, complex cells could respond to a stimulus within RF regardless of its position, which meant complex cells weren't selective to stimulus orientation. To explain the structure of complex cell RF, they proposed a hierarchical model in which simple cells with similar orientation preferences but different RF positions (also known as subunits) fed into a complex cell [4].

Since then, various models of complex cells have been proposed using the spatial relationships among subunit RFs. The most popular model among them is the energy model. The first energy model was proposed by Green and Swets in 1966, which consisted of two linear band-pass filters with quadrature phase followed by a squaring device [29]. As a model for binocular complex cells, in 1990 Ohzawa *et al* demonstrated a binocular version of the energy

model which could respond to the stimulus energy associated with binocular interactions as shown in Figure 2.3 [30]. The response of subunit S1 is given by [4]:

$$R_{S1}(X_L, X_R) = Pos[exp(-kX_L^2)cos(2\pi f X_L) + exp(-kX_R^2)cos(2\pi f X_R + \psi)]^2 \tag{2.13}$$

where k is the factor that determines the width of the subunit RFs and f represents spatial frequency which are assumed to be same for both eyes. ψ is the phase difference between the left and right monocular RFs. $Pos[y]$ is a half-rectifying function. So by adding the outputs of S1, S2, S3 and S4 together, the response of complex cell C1 can be obtained [30]:

$$R_{C1}(X_L, X_R) = R_{S1} + R_{S2} + R_{S3} + R_{S4}$$

$$= \{exp(-kX_L^2)cos(2\pi f X_L) + exp(-kX_R^2)cos(2\pi f X_R + \psi)\}^2$$

$$+ \{exp(-kX_L^2)cos(2\pi f X_L) + exp(-kX_R^2)cos(2\pi f X_R + \psi)\}^2 \tag{2.14}$$

To gain a better understanding upon complex cells, Ohzawa *et al* recorded from total 257 neurons in the striate cortex of 18 adult cats, and among them, 115 cells are classified as complex on the basis of subjective criteria [12]. The binocular interaction RFs they recorded were decomposed into functional subunits with the singular value decoposition (SVD). According to their results, the first two SVD components were consistent with subunits of an energy model with respect to RF properties [4].

The responses for two matched contrast sign condition (dark-dark and bright-bright) are almost identical with a diagonal elongated excitation region. For opposite contrast situations (dark-bright and bright-dark), there are two excitation regions located on both sides of the diagonal line. To test the performance of their hierarchical energy model, Ohzawa *et al* compared the contour plot generated by Equation 2.14 with the binocular interaction RFs they measured. Their results indicates that the responses of complex cells are fit well by a single unit energy model. However, for some complex cells, the energy model with a single unit did not give satisfaction fits [4].

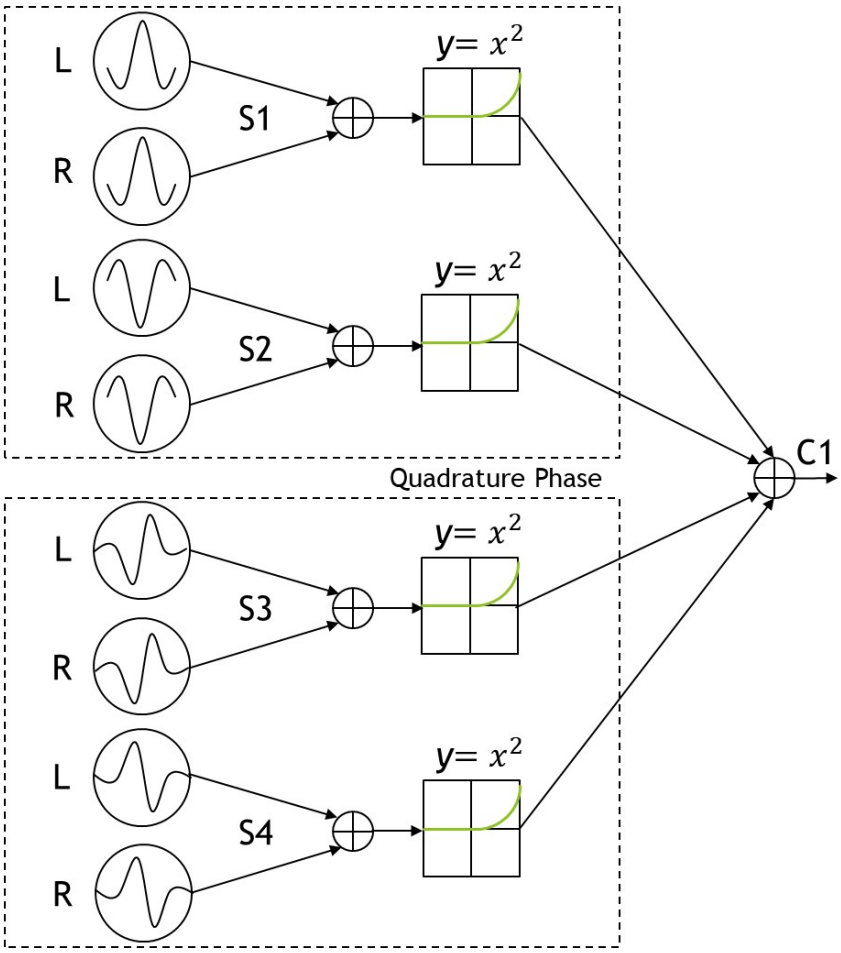

Figure 2.3. Structure of binocular disparity energy model. The energy model of a complex cell (C1) tuned to 0 disparity consists of two major units with quadrature phase (four subunits S) [12]. S1-S4 represent the four simple cell-like subunits whose monocular RFs are modeled by one-dimensional Gabor filters. Spatial phases of monocular RFs in S1 and S3 are 90 degrees apart (same with S2 and S4) which enables the binocular energy model can respond to a stimulus despite its spatial phase. Each subunit combines input from left eye and right eye linearly and output of each subunit is followed by a half-squaring nonlinearity, representing the fact that "only postsynaptic potentials that exceed a threshold value elicit action potentials" [12]. Redrawn from [12] (Fig.8).

.

As we illustrated before, simple cells can calculate the cross-correlation of the left and right eye images which were bandpass filtered because of the multiplicative binocular interaction. Because complex cells subunits are functionally similar to simple cells, complex cells can also perform something analogous to an interocular cross-correlation in a local region [4]. Although both simple cells and complex cells can calculate interocular cross-correlation which is believed to be a fundamental computation for stereopsis, simple cells responses depend on monocular stimulus phases whereas complex cells respond to stimulus regardless its phase.

2.2.3. RF Position and Phase Disparities of Complex-Cell Subunits

We already showed that simple cells could respond to binocular stimulus. But how do they encode binocular disparity? There are two possible hypotheses for this question (Figure 2.4):

1. Position shift: The left and right eye RFs of a neuron have the same spatial structure but with different positions on retina. Through the RF position disparity, binocular disparity can be encoded.

2. Phase shift: The left and right eye RFs of a neuron have the same position on retina but with different RF phases (profiles). In this case, binocular disparity is encoded by phase disparity.

The left and right eye RFs of a simple cell in Position shift model are given by following equations [4]:

$$g_l^{pos}(x,y) = g(x,y,\phi)$$

$$g_r^{pos}(x,y) = g(x+d,y,\phi) \tag{2.15}$$

where g(x,y) is a Gabor function (Equation 2.12) and d represents the position shift between monocular RFs. In contrast, in phase shift model monocular RFs are expressed as [4]:

$$g_l^{pha}(x,y) = g(x,y,\phi)$$

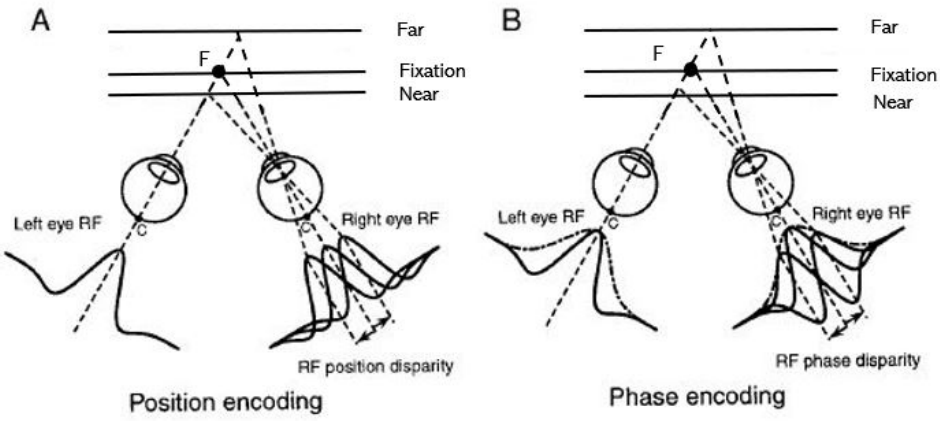

Figure 2.4. Two hypotheses of how cortical cells encode binocular disparity. A. Position encoding. Binocular disparity is encoded by a position difference between the left and right eye RFs which have the same spatial profile. B. Phase encoding. Binocular disparity is encoded by a phase difference between the left and right eye RFs which have the same position. Figure adapted from [13] (Fig.1).

.

$$g_r^{pha}(x,y) = g(x,y,\phi+\Delta\phi) \tag{2.16}$$

where $\Delta\phi$ is the phase shift between monocular RFs. Using these equations, we can derive complex cells responses based on the energy model [28]:

$$r_q^{pos} \approx 4A^2cos^2(\frac{\omega(D-d)}{2}) + \frac{D-d}{\sigma_x}4ABcos(\frac{\omega(D-d)}{2})cos(\alpha-\beta+\frac{\omega(D-d)}{2}) \tag{2.17}$$

$$r_q^{pha} \approx 4A^2cos^2(\frac{\omega D-\Delta\phi}{2}) + \frac{D}{\sigma_x}4ABcos(\frac{\omega D-\Delta\phi}{2})cos(\alpha-\beta+\frac{\omega D-\Delta\phi}{2}) \tag{2.18}$$

for the position shift and phase shift models. D is the disparity of a stereo image pair. A and α are the local Fourier amplitude and phase (evaluated at the RF's preferred frequency) of the stimulus patch filtered by the RF Gaussian envelope. B and β represent the similar amplitude and phase of the stimulus patch filtered by "the first-order derivative of the RF Gaussian envelope" [28]. Compared with amplitudes A and B, the phases α and β are more dependent on the luminance distribution of the stereo images. And the preference disparity

of a complex cell based on position shift and phase shift models are [28]:

$$D_{pref}^{pos} \approx d \tag{2.19}$$

$$D_{pref}^{pha} \approx \frac{\Delta\phi}{\omega} \tag{2.20}$$

According to Anzai *et al*,both position shift and phase shift mechanisms have been found in the cats striate cortex [13]. However, based on their experiments Anzai *et al* illustrated that position disparities were generally small, in other words, were only suitable for encoding small binocular disparities. More importantly, position disparities didn't show any correlation with orientation or spatial frequency preference [13]. They argued that position shift might be a "byproduct of random jitter in RF position" [13]. To the contrary, phase disparities covered a wider binocular disparities range and showed an orientation anisotropy, which were generally within the quarter cycle limit and provided a basis for the size-disparity correlation observed in psychophysics [13]. So Anzai *et al* concluded that binocular disparities were mainly encoded by phase shift mechanism, meanwhile, position shift mechanism could still play an important role in high spatial frequency cases in which phase disparity becomes significantly small.

The energy model is a well defined computational model which provides a good description of the first stages of the cortical binocular processing. But in energy model, no scanning eye movements are considered, as it assumes that that parallel visual processing continues through to the highest level. As we argued before, we believe scanning eye movements are necessary for stereopsis. Can we take eye movements into account? This question is the second motivation for this thesis.

Chapter 3

Computational Theories for Human Stereo Vision

3.1. Dynamic Stereo Matching Algorithm

Does the primary visual cortex have similar structures to the Jeffress model which are responsible for binocular disparity encoding? Can we take eye movements into account for computational stereo matching? Driven by the two questions, we proposed a dynamic stereo matching algorithm which consists of a hierarchical structure as shown in Figure 3.1.

The first layer represents the low-level visual system which translates spatial luminance information into continuous-time signals with scanning eye movements. Notice that, We assume the scanning eye movements are along the x-axis. And the output S of the first layer at given time t can be expressed as:

$$S(t) = I(x_0 + \eta t, y_0) \tag{3.1}$$

where $I(x, y)$ is the luminance at position (x, y) and (x_0, y_0) is the initial position of the monocular RFs. η represents the moving speed of the scanning eye movements.

The second layer is a time-delay neural network which is similar to the Jeffress model. Coincidence detectors tuned to different disparities ($-d$ to d) are used to encode binocular disparities. The coincidence detector output activates only when it receives same signals within a time window accepted as at the same time and in parallel at both inputs which is given by following equation:

$$C_d = \begin{cases} 1 & S_L(t + \frac{d}{\eta}) = S_R(t) \quad within \quad \Delta t \\ 0 & Otherwise \end{cases} \tag{3.2}$$

d is the preference disparity. S_L and S_R are the left and right inputs from the first layer. Δt is the time window of coincidence detector.

The third layer is the integrate-and-fire (IF) neuron layer. Each IF neuron is connected to a coincidence detector and integrates its input within a time window. The function of an IF neuron is:

$$Y = f(\int_{\Delta T} C_d \, dt \ ) \tag{3.3}$$

where $f(u)$ is nonlinear activation functions. ΔT is the time window of IF neurons. Notice that, the inhibitory connections between IF neurons ensure only one IF neuron will be firing (winner-take-all).

Random-dot stereograms (RDSs) are used to test the performance of our algorithm, which were invented by Julesz in 1971 [31]. RDSs consist of random dots when viewed monocularly but fuse when viewed stereoscopically to yield patterns separated in depth (Figure 3.2). More specifically, the monocular images are arranged identically, except that a portion of the dots is moved to the left or the right in one of the images to create either a crossed or an uncrossed disparity. This creates the experience that part of the image is either in front of or behind the rest of the dots. To have a better estimation upon our algorithm, we introduce the difference map which shows the differences between the true disparity map and the estimated disparity map.

The output of our algorithm is given by Figure 3.3. In this example, the depth pattern embedded in the RDS contains two different disparities (0 and +3). Figure 3.4 shows another example with two disparities embedded. From the difference maps, we can find that the accuracy of our model is relatively satisfying. In our simulation, the nonlinear activation function of the IF neuron is:

$$f(u) = \begin{cases} 1 & u > \theta \\ 0 & Otherwise \end{cases} \tag{3.4}$$

where θ is the threshold.

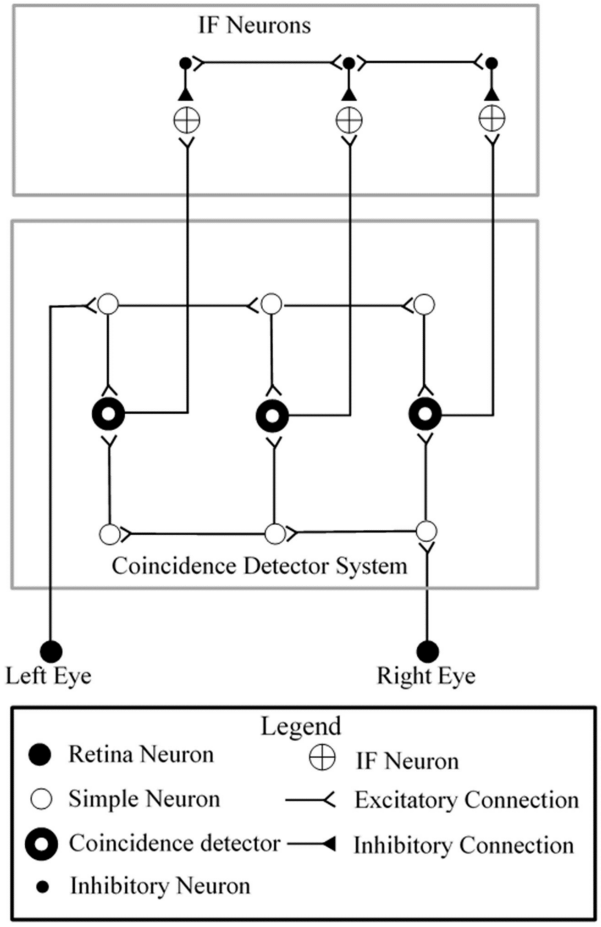

Figure 3.1. Structure of the dynamic stereo matching algorithm. With scanning eye movements monocular RFs transform spatial information into continuous-time signals. The second layer contains a time-delay neural network similar to the Jeffress model. In this case, the coincidence detector system is tuned to three different disparities from -1 to 1.

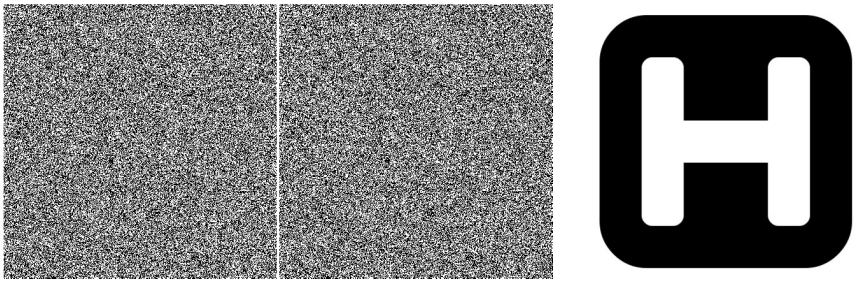

Figure 3.2. An example of RDS and the embedded pattern

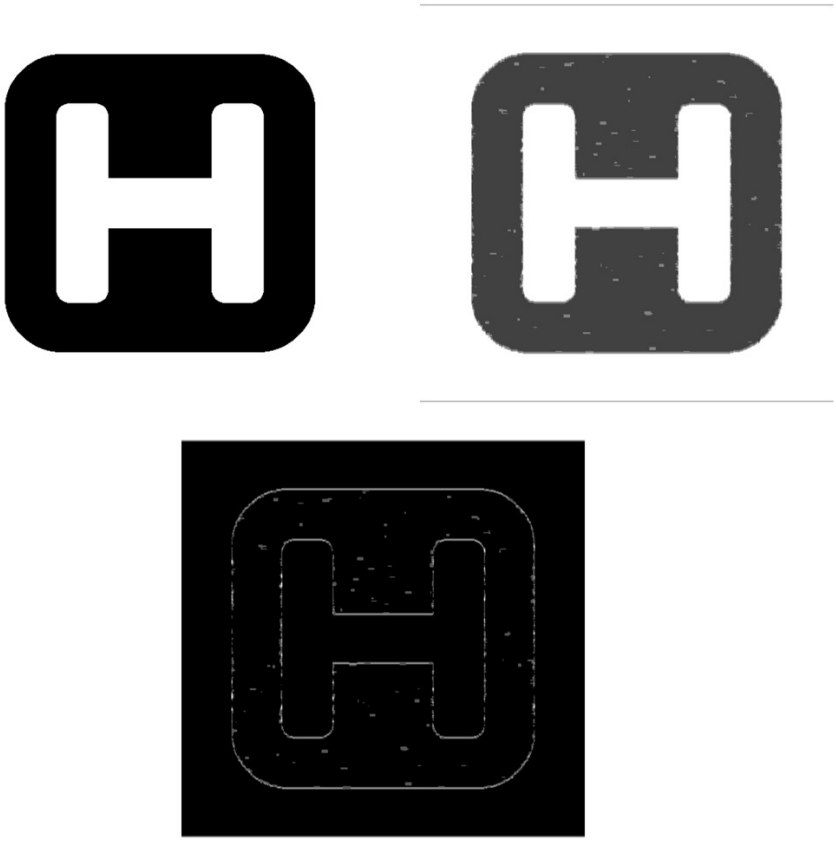

Figure 3.3. Output of the dynamic stereo matching algorithm. The difference map (second row) shows the differences between the true disparity map (first row left side) and the estimated disparity map (first row right side). In this case, the coincidence detector system is tuned to seven different disparities from -3 to 3.

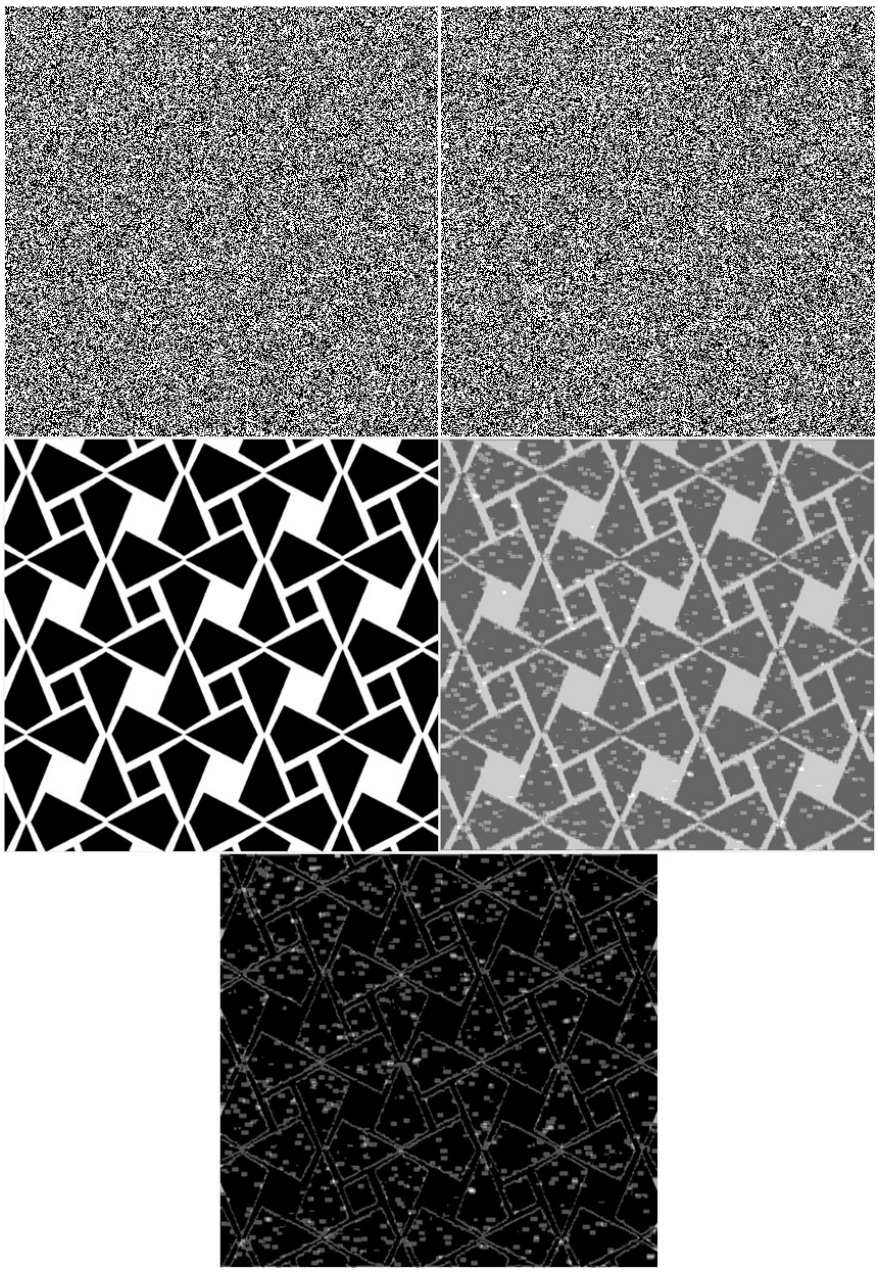

Figure 3.4. The first row is a pair of RDS. The second row shows the pattern embedded in the RDS (left) and the output disparity map of the dynamic stereo matching algorithm (right). The third row is the difference map. In this example, the coincidence detector system is tuned to five different disparities from -2 to 2.

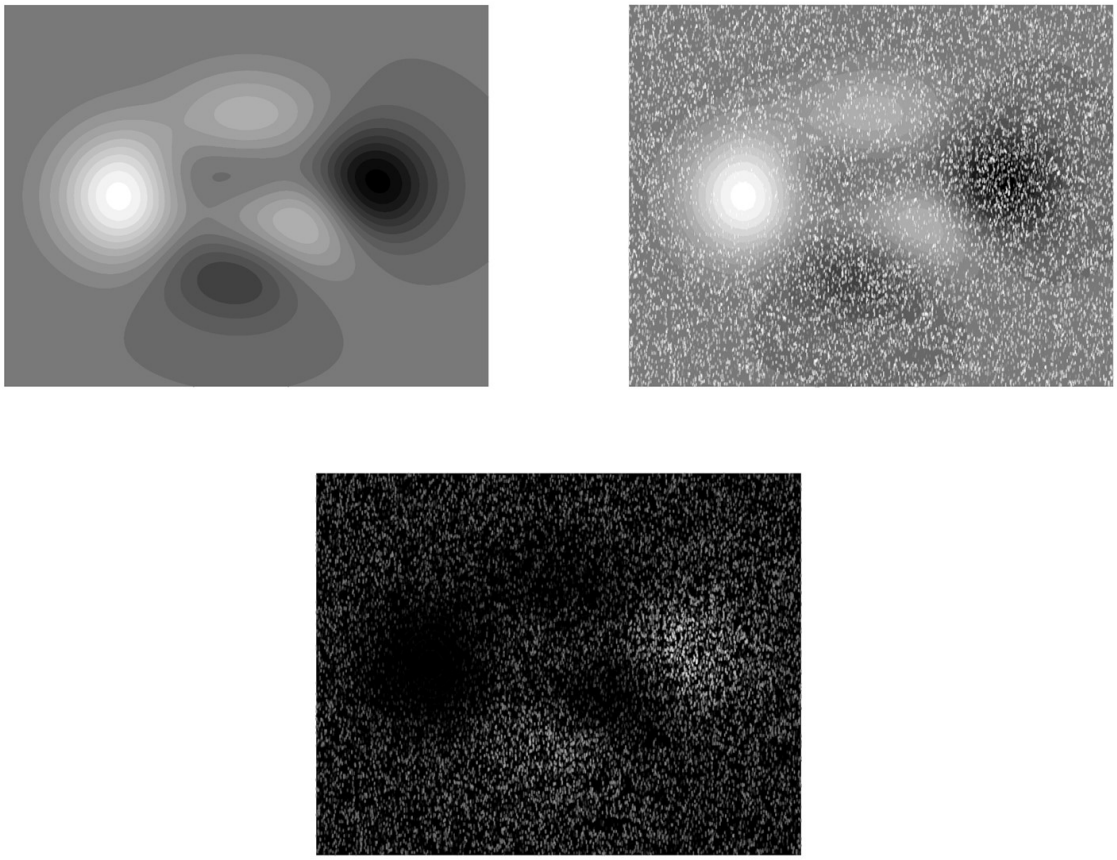

Figure 3.5. The disparity ground truth (first row left side) and the outcome of the dynamic model (first row right side) (20 disparities from -9 to 10). There are lots of false matches as shown in the difference map (second row).

Further, we tried more complex RDSs with more disparity levels embedded. The outcome is shown in Figure 3.5. Obviously, when the number of disparities goes up, the performance of our algorithm becomes worse. The possible explanation for it is that in our model the coincidence detector only takes information only from position (x, y) (a single pixel) which may introduce many false matches. Therefore, we need to apply a new mechanism to extract signals from a local part of the surface which can enhance the correlation between the left and right inputs.

3.1.1. Dynamic Stereo Matching Algorithm with Gabor Filters

The majority of neurons in primary visual cortex exhibit orientation selectivity [3, 9, 26]. Generally, these orientation-selective neurons with center-surround RFs are sensitive to the luminance changes (contrast) on the surface. Both orientation selectivity and contrast sensitivity can be well described by a two-dimensional Gabor filter. In 1987, J. P. Jones and L. A. Palmer illustrated in their paper that most of the simple cells' binocular RFs are Cartesian separable and the separated sections along the width and length axis are similar to one-dimensional Gabor filters [14] (Figure 3.6).

To enhance the performance of the dynamic stereo matching algorithm, we introduce one-dimensional Gabor filters to the first layer which are used to model monocular RFs. For simplicity, we assume that the corresponding monocular RFs in the left and right eye have the same spatial profile, phase, and relative position on the retina. Then the output signal of the low-level visual system layer becomes:

$$S_{new}(t) = g(x) * I(x, y_n) = \int_{\Delta x} g(x^{'}) I(x - x^{'}, y_n)\, dx^{'} \tag{3.5}$$

where $g(x)$ is a one-dimensional Gabor filter and Δx is the length of the one-dimensional Gabor filter. $I(x, y_n)$ represents the luminance of the nth row form the monocular stereo image. $*$ denotes convolution. And location x at the given time t is given by:

$$x = x_0 + \eta t \tag{3.6}$$

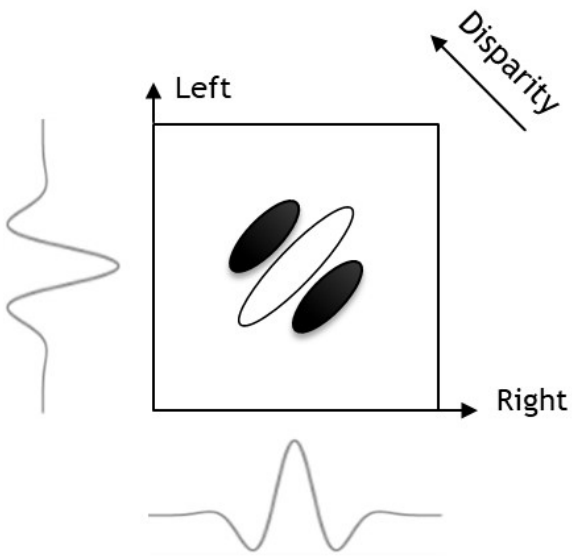

Figure 3.6. The Cartcsian separation of a simple cell's binocular RF. The bright ellipse is the excitatory subregion and the dark ellipses represent the inhibitory subregions. The sections along the left axis and right axis are fit well by Gabor filters [14].

where x_0 is the initial position and η is the scanning speed of the left and right eye.

With the one-dimensional Gabor filter, we can extract information not only from the pixel located at (x, y) but also from its surrounding pixels (in range Δx along x direction). And we use sigmoid as the activation function of IF neurons which is expressed as:

$$f(u) = \frac{1}{1 + e^{-x}} \tag{3.7}$$

The output of this model for the same RDS with 20 disparities is shown in Figure 3.7. Compared with Figure 3.5, we can conclude that one-dimensional Gabor filters can eliminate false matches effectively. However, our test results indicate that our model is not compatible with natural stereo images in which the luminance of the corresponding points is slightly different. In other words, when the difference between two corresponding points is relatively large, our model works poorly because of the poor robustness of the coincidence detector function.

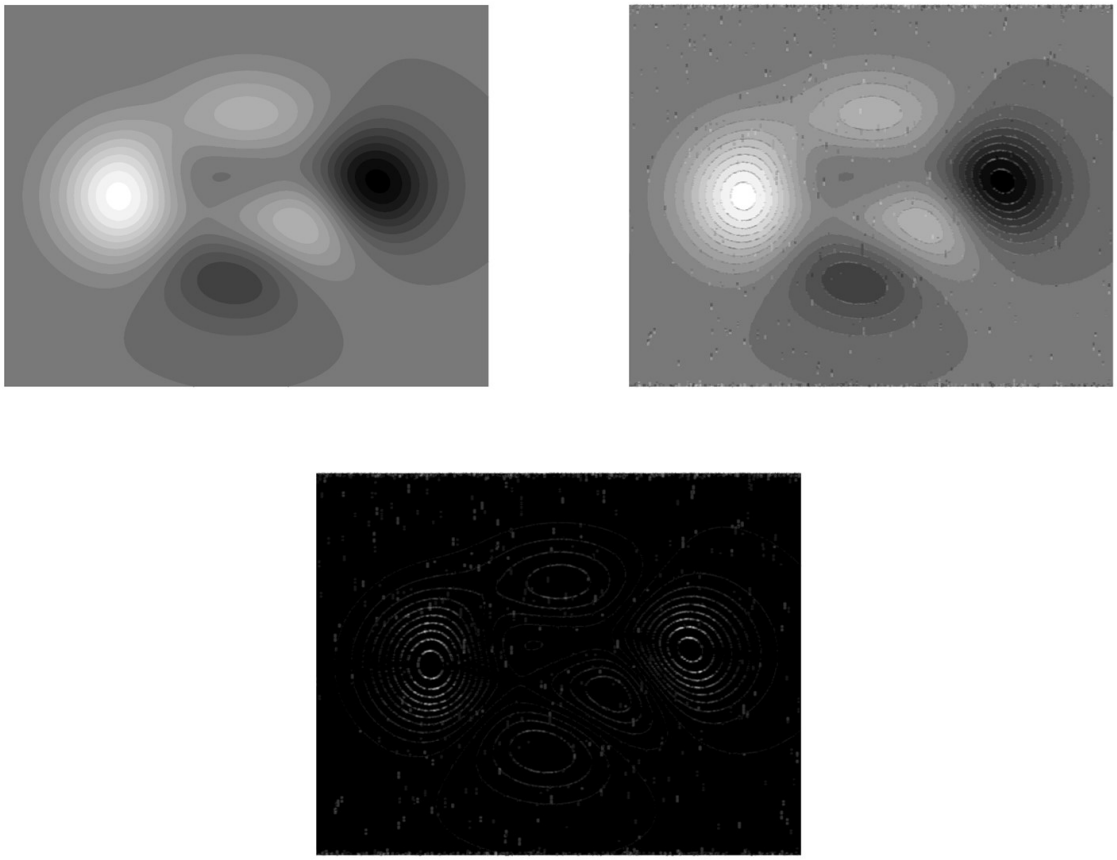

Figure 3.7. The disparity ground truth and the outcome of the dynamic model with 2D Gabor filters (20 disparities from -9 to 10). Compared with Figure 3.5, the number of the false matches has been reduced significantly.

Figure 3.8. A pair of natural stereo images (first row) and the disparity map recovered by the dynamic algorithm with Gabor filters (second row). In this case, the function of the second layer and the third layer is similar to the cross-correlation function (Equation 3.11).

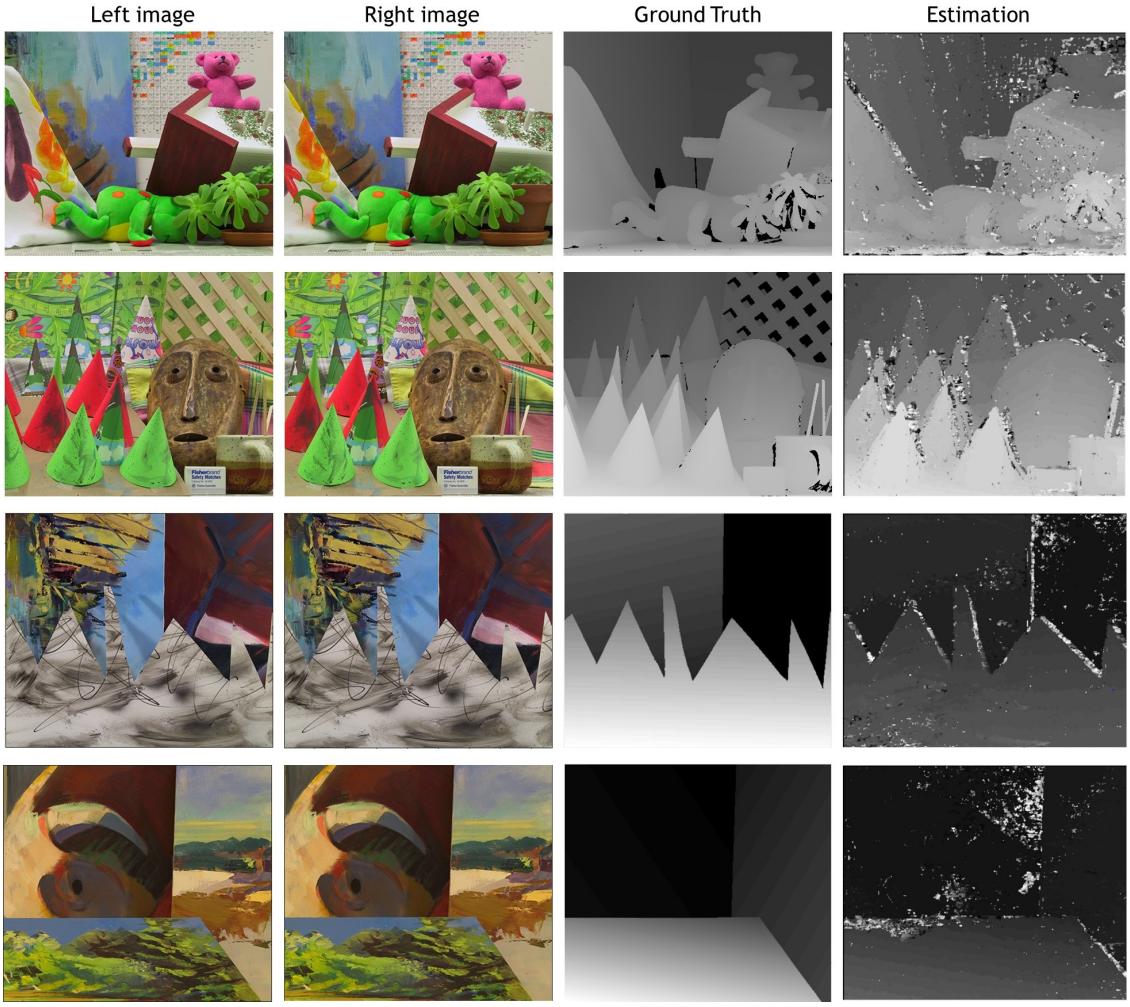

Figure 3.9. Natural stereo images from the Middlebury database (first two columns). In this case, the function of the second layer and the third layer is similar to the cross-correlation function (Equation 3.11). Comparing the estimated disparity maps (fourth column) with the true disparity maps (third column), we can find that the overall acuity is relatively low. There are lots of false matches in the estimations, especially for the areas around the edges.

As we illustrated in the former part, both simple cells and complex cells are found to be able to calculate interocular cross-correlation. So can we make our algorithm compatible with the natural stereo images by applying cross-correlation or similar mechanism ? The normalized cross-correlation of the left and right eye inputs is given by following equation:

$$NC = \frac{\sum_{i=1}^{N}(S_{li}(t+dt) - \bar{S}_l)(S_{ri}(t) - \bar{S}_r)}{\sqrt{\sum_{i=1}^{N}(S_{li}(t+dt) - \bar{S}_l)^2 \sum_{i=1}^{N}(S_{ri}(t) - \bar{S}_r)^2}} \tag{3.8}$$

where S_l and S_r are the inputs from the left and right eye and N denotes the length of the time window Δt. dt is the corresponding time shift between two input signal which is proportional to the disparity. $\bar{S}_l$ and $\bar{S}_r$ are the means of the two inputs. Based on this equation, we assume that the coincidence detector is able to calculate a weighted multiplication given by:

$$C_d = \frac{(S_l(t+\frac{b}{\eta}) - \bar{S}_l)}{\sqrt{\sum_{i=1}^{N_t}(S_{li}(t+\frac{b}{\eta}) - \bar{S}_l)^2}} \cdot \frac{(S_r(t) - \bar{S}_r)}{\sqrt{\sum_{i=1}^{N_t}(S_{ri}(t) - \bar{S}_r)^2}} \tag{3.9}$$

where N_t is the length of the Δt. Then the output of the corresponding IF neuron is expressed as:

$$O_{IF} = \int_{\Delta T} C_d \, dt \tag{3.10}$$

if we plug in Equation 3.9 in to it:

$$O_{IF} = \int_{\Delta T} \frac{(S_l(t+\frac{b}{\eta}) - \bar{S}_l)(S_r(t) - \bar{S}_r)}{\sqrt{\sum_{i=1}^{N_t}(S_{li}(t+\frac{b}{\eta}) - \bar{S}_l)^2 \sum_{i=1}^{N_t}(S_{ri}(t) - \bar{S}_r)^2}} \, dt \tag{3.11}$$

which is similar to the cross-correlation function. Winner-take-all algorithm is used to determine the corresponding binocular disparity D between inputs S_l and S_r which is given by the following equation:

$$D = d_{\max(O_{IF_d})} \tag{3.12}$$

As shown in Figure 3.8, with the new function, our dynamic model is able to recover a rough disparity map from the natural stereo image. To evaluate the performance of our algorithm, we use the natural stereo images from the Middlebury database (Figure 3.9). Obviously, the overall acuity of the estimated disparity maps is relatively low. More importantly, the number of disparities that our algorithm can handle is very limited compared with the true disparity number because of the second and third layers' function. That is why the same surface may not necessarily have the same disparity value in the true disparity maps and the estimations. And it makes hard for us to calculate the difference maps like we did for RDSs. Notice that, the areas around the edges contain more false matches than others. This phenomenon might be the result of the Gabor filters' poor responses for the objects' edges.

3.2. Combination of Receptive Fields (CORF) Model Based Dynamic Stereo Matching Algorithm

The combination of Receptive Fields (CORF) Model is a computational model which takes afferent inputs from LGN cells with center-surround RFs. It was proposed by G. Azzopardi and N. Petkov in 2012 [5]. The structure of the CORF computational model is presented in Figure 3.10.

Each of the light and dark disks in the figure represents the RF of a subunit which receives input from a pool of center-on or center-off LGN cells which are sensitive to contrast. A subunit computes the weighted summation of the outputs of LGN cells. And the orientation selectivity of the CORF model is achieved by combining the responses of given subunits with appropriate polarities and alignment of their RFs [5]. In the CORF model, the LGN cell with a center-on RF is modeled by a difference of 2D Gaussian functions [5]:

$$DoG_{\sigma}^{+} = \frac{1}{2\pi\sigma^2} exp(-\frac{x^2+y^2}{2\sigma^2}) - \frac{1}{2\pi(0.5\sigma)^2} exp(-\frac{x^2+y^2}{2(0.5\sigma)^2}) \tag{3.13}$$

where σ denotes the standard deviation of the outer Gaussian function and the standard deviation of the inner Gaussian function is 0.5σ. This is in accordance with the LGN

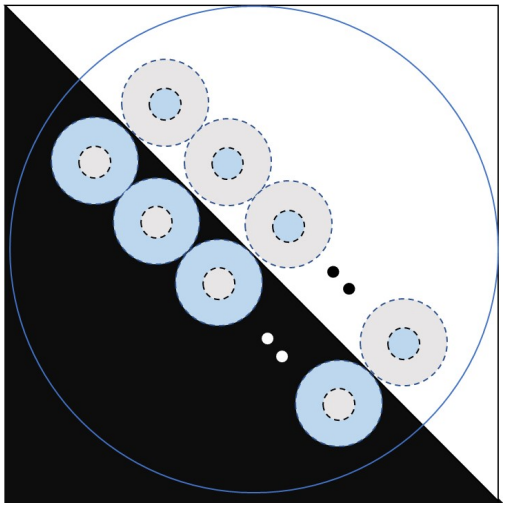

Figure 3.10. Structure of the CORF computational model of a simple cell which consists of two subunits: center-on unit (right) and center-off unit (left) [5]. By combining the responses of two parallel subunits, orientation selectivity of a simple cell is achieved. Each center-surround subunit calculates a weighted summation of the responses of a group of local LGN cells [5]. Redrawn from [5].

cells found in mammals [5]. The LGN cell with a center-off RF is given by the following equation [5]:

$$DoG_{\sigma}^{-} = -DoG_{\sigma}^{+} \tag{3.14}$$

The response of a model LGN cell with a RF centered at image coordinates (x, y) is computed by linear spatial summation of the intensity distribution $I(x', y')$ in the input image which is weighted by the function $DoG(x - x', y - y')$ [5]. The weighted summation is followed by half-wave rectification [5]:

$$c_{\sigma}^{\delta}(x, y) = |I * DoG_{\sigma}^{\delta}| \tag{3.15}$$

where δ represents the polarity (+ for center-on and - for center-off) of the DoG function.

The response of the subunit which performs a weighted linear summation is expressed as [5]:

$$S_{\delta_i,\sigma_i,\rho_i,\phi_i}(x,y) = \sum_{x'}\sum_{y'}\{c^{\delta_i}_{\sigma_i}(x-\Delta x_i - x', y-\Delta y_i - y')G_{\sigma'}(x',y')\} \tag{3.16}$$

$$-3\sigma' \leq x', \quad y' \leq 3\sigma'$$

where $\Delta x_i = -\rho_i cos\phi_i$ and $\Delta y_i = -\rho_i sin\phi_i$. G'_σ is the two-dimensional Gaussian weighting function for the CORF model cell with the standard deviation σ' [5]. Equation 3.16 is a convolution of the weighting function G'_σ with the function $c^{\delta_i}_{\sigma_i}$ which is shifted by $(\Delta x_i, \Delta y_i)$.

The response of a CORF model cell is given by the following equation [5]:

$$r_s(x,y) = (\prod_{i=1}^{|S|}(S_{\delta_i,\sigma_i,\rho_i,\phi_i})^{\omega_i})^{\frac{1}{\sum_{i=1}^{|S|}\omega_i}} \tag{3.17}$$

where $w_i = e^{-\frac{\rho_i^2}{2\sigma'^2}}$ and $\sigma' = \frac{1}{3}\max_{i\in\{1\ldots|S|\}}\rho_i$. Computing the product of subunit responses indicates that the concerned CORF model cell is activated only when all its afferent subunits are active [5].

The orientation selectivity of a CORF model cell depends on the orientation of the edge used for the model configuration. The orientation-selective response is defined as follows [5]:

$$\Re_\psi(S) = \{(\delta_i, \sigma_i, \rho_i, \phi_i + \psi) | \forall(\delta_i, \sigma_i, \rho_i, \phi_i) \in S\} \tag{3.18}$$

To detect contours of any orientation, we can merge the responses of CORF model cells with different orientation selectivities by taking the maximum value at a given location (x,y) [5]:

$$\hat{r}_s = \max_{\psi\in\Psi}\{r_\Re(S)(x,y)\} \tag{3.19}$$

where Ψ is a set of n_θ equidistant orientations. Based on their experiments, the choice of $n_\theta = 12$ ensures sufficient response for all orientations [5]. The authors argued that compared with Gabor filters, the CORF model they proposed was more robust to noise and

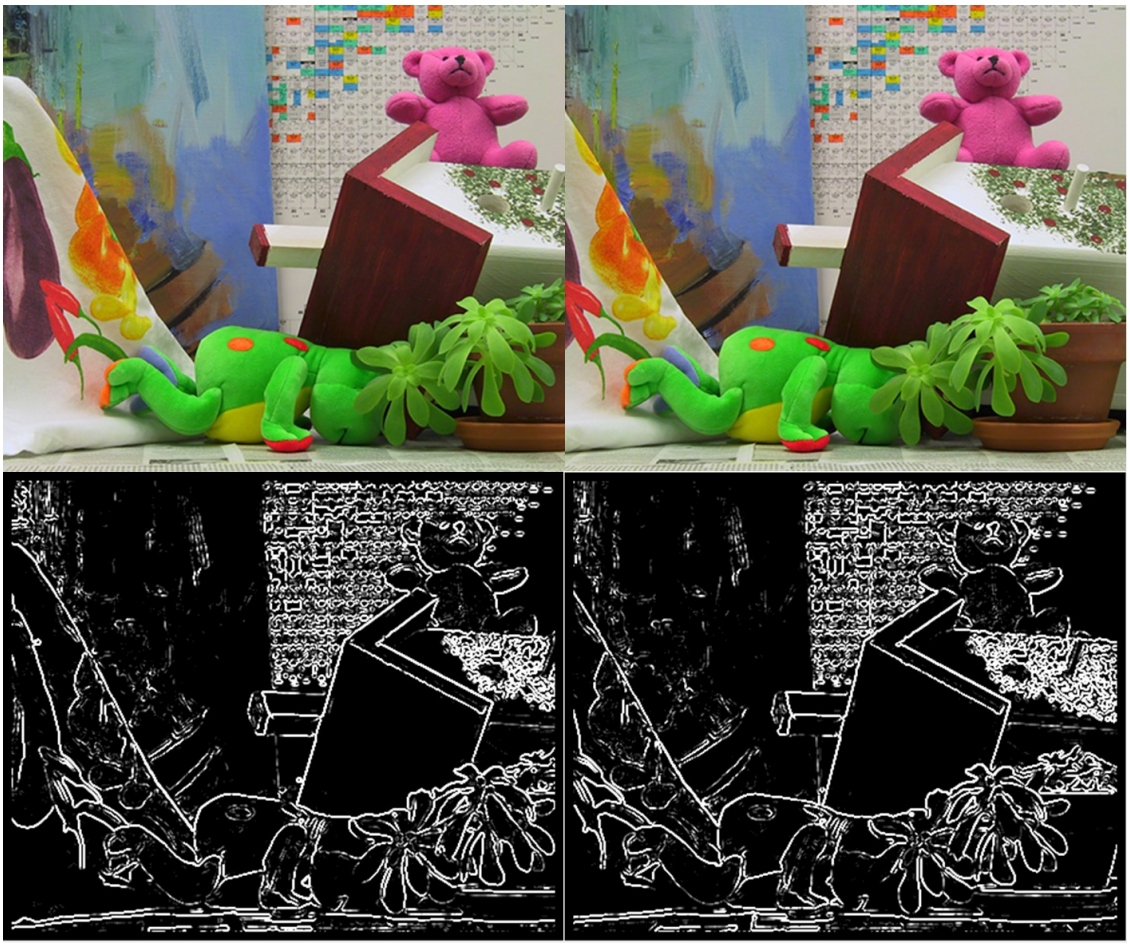

Figure 3.11. The output of the CORF model with 24 angles. The CORF model's responses are consistent with the orientation selectivity and contrast sensitivity (edge preference) of the simple cells.

had better edge localization [5]. Are we able to take advantages of the CORF model to enhance the disparity acuity? The flow chart of the dynamic model with the CORF model embedded is shown in Figure 3.12.

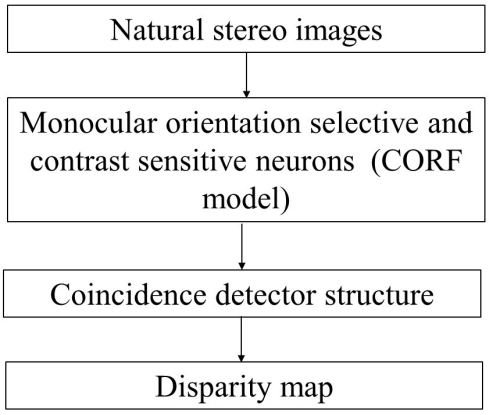

Figure 3.12. The flow chart of the dynamic algorithm with the CORF model.

The output disparity map is shown in Figure 3.13. Obviously, the overall stereoscopic acuity becomes worse compared with the Gabor filters embedded algorithm. Based on our analysis, the best explanation for this phenomenon is that the CORF model is sensitive to the edges but has almost no response to the smooth surface (Figure 3.12), which will introduce many false matches to the final outcomes. In other words, the CORF model's response for a given point located on a smooth surface is too small to find the correct corresponding point form another image, which will also harm the stereo analysis on edges.

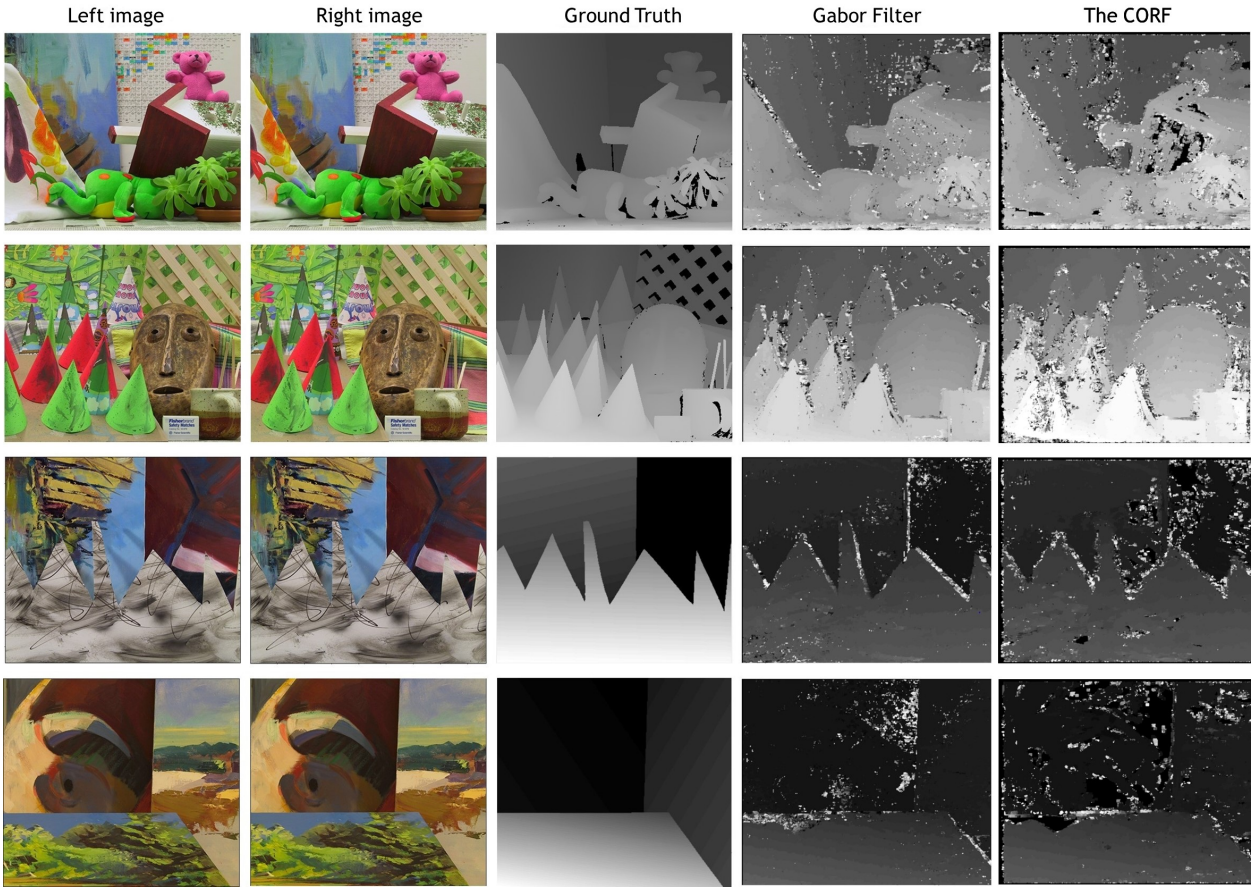

Figure 3.13. The estimated disparity maps of the dynamic stereo matching algorithm with the CORF model (fifth column). Compared with the estimations of the algorithm with Gabor filters (fourth column), the overall acuity becomes worse. The possible cause is that the CORF model introduces uncorrelated features to the stereo images.

Chapter 4

Discussion

In this thesis, we explored the feasibility of applying the Jeffress model in stereo matching. Compared with the models we talked about in chapter two, the main difference of our dynamic model is that scanning eye movements are considered. The simulation results indicate that the time-delay neural network is a possible solution for stereopsis. Although lots of effort has been made to reveal the underlying role of the eye movements in stereo matching, the final answer is still there waiting for someone to discover. We believe that our model gives a good insight into this question that only with eye movements we are able to generate the monocular inputs with time differences.

Compared with the energy model, our algorithms explain the disparity selectivity of the complex cell via a time-delay neural network. A potential advantage of our algorithms compared with the energy model is lower computational complexity, as our algorithms process the spatial information temporally.

One of the biggest concerns we have for our stereo matching algorithm is biological feasibility. Mechanisms of the coincidence detector and IF neuron have already been well studied [32]. In the Jeffress model, the disparity is encoded by the time difference between two inputs. Our assumption that the primary visual cortex has the similar structure to the Jeffress model is supported by the long-range horizontal connections found in V1. These connections enable cells to integrate information from a large part of the binocular visual field. As we illustrate in the previous chapter, the cat brain has been found to be able to perform the calculation which is similar to the cross-correlation [3, 4, 13]. Notice that in Equation 3.11, if the length of ΔT equals to N_t, this equation becomes the normalized cross-correlation. And if we take out the weighting factors which might not be biologically realistic, the outputs of IF neurons are similar to the cross-correlation. Based on the outputs

of IF neurons, we can estimate the binocular disparities. So we believe that our algorithm is a possible explanation for stereopsis.

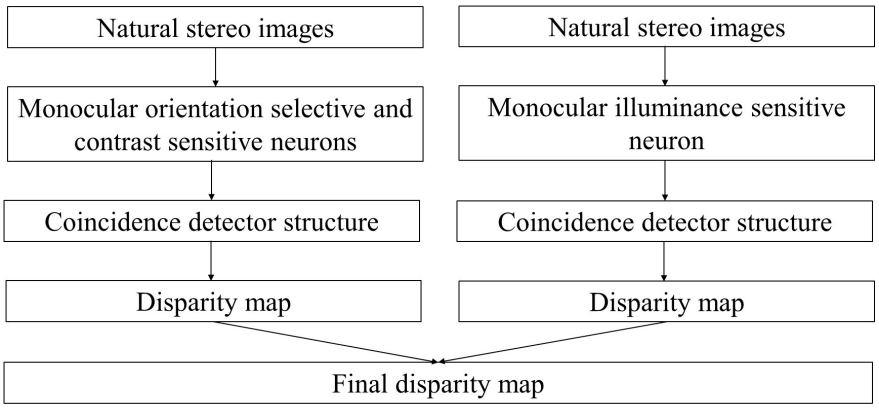

Figure 4.1. The structure of a possible stereo matching algorithm.

We also notice that both Gabor filters and the CORF model are specialized to respond to contours. As we illustrated in chapter one, besides the edge preference cells there are cells are specialized to respond to surface other than edges. Is it possible that both kinds of cells are essential for stereopsis? Can we come up an algorithm similar to the one illustrated in Figure 4.1?

More importantly, compared with the current stereo matching models [33,34], the stereoscopic acuity of our algorithm is still far from satisfying. Our job is only the beginning. More problems and challenges are waiting for us. As our vision is so vivid and accurate, we believe that studies on the visual systems of animals can help us enhance the stereoscopic acuity greatly.

BIBLIOGRAPHY

[1] D. D. Stettler, A. Das, J. Bennett, and C. D. Gilbert, "Lateral connectivity and contextual interactions in macaque primary visual cortex," *Neuron*, vol. 36, no. 4, pp. 739–750, 2002.

[2] L. A. Jeffress, "A place theory of sound localization.," *Journal of comparative and physiological psychology*, vol. 41, no. 1, p. 35, 1948.

[3] A. Anzai, I. Ohzawa, and R. D. Freeman, "Neural mechanisms for processing binocular information i. simple cells," *Journal of neurophysiology*, vol. 82, no. 2, pp. 891–908, 1999.

[4] A. Anzai, I. Ohzawa, and R. D. Freeman, "Neural mechanisms for processing binocular information ii. complex cells," *Journal of Neurophysiology*, vol. 82, no. 2, pp. 909–924, 1999.

[5] G. Azzopardi and N. Petkov, "A corf computational model of a simple cell that relies on lgn input outperforms the gabor function model," *Biological cybernetics*, vol. 106, no. 3, pp. 177–189, 2012.

[6] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proc. R. Soc. Lond. B*, vol. 204, no. 1156, pp. 301–328, 1979.

[7] A. L. Yarbus, "Eye movements during perception of complex objects," in *Eye movements and vision*, pp. 171–211, Springer, 1967.

[8] D. Purves, G. J. Augustine, D. Fitzpatrick, L. C. Katz, A.-S. LaMantia, J. O. McNamara, S. M. Williams, *et al.*, "Types of eye movements and their functions," *Neuroscience*, pp. 361–390, 2001.

[9] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, A. J. Hudspeth, *et al.*, *Principles of neural science*, vol. 4. McGraw-hill New York, 2000.

[10] W. H. Bosking, Y. Zhang, B. Schofield, and D. Fitzpatrick, "Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex," *Journal of neuroscience*, vol. 17, no. 6, pp. 2112–2127, 1997.

[11] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, no. 4262, pp. 283–287, 1976.

[12] I. Ohzawa, G. C. DeANGELIS, and R. D. Freeman, "Encoding of binocular disparity by complex cells in the cat's visual cortex," *Journal of neurophysiology*, vol. 77, no. 6, pp. 2879–2909, 1997.

[13] A. Anzai, I. Ohzawa, and R. D. Freeman, "Neural mechanisms for encoding binocular disparity: receptive field position versus phase," *Journal of Neurophysiology*, vol. 82, no. 2, pp. 874–890, 1999.

[14] J. P. Jones and L. A. Palmer, "The two-dimensional spatial structure of simple receptive fields in cat striate cortex," *Journal of neurophysiology*, vol. 58, no. 6, pp. 1187–1211, 1987.

[15] D. Vishwanath, "Toward a new theory of stereopsis.," *Psychological review*, vol. 121, no. 2, p. 151, 2014.

[16] N. J. Wade, H. Ono, and L. Lillakas, "Leonardo da vinci's struggles with representations of reality," *Leonardo*, vol. 34, no. 3, pp. 231–235, 2001.

[17] C. Wheatstone, "On some remarkable and hitherto unobserved phenomena of binocular vision.," *The Optometric weekly*, vol. 53, p. 2311, 1962.

[18] K. Rayner, "Eye movements and visual cognition: Introduction," in *Eye Movements and Visual Cognition*, pp. 1–7, Springer, 1992.

[19] B. Clark, "An eye-movement study of stereoscopic vision," *The American Journal of Psychology*, vol. 48, no. 1, pp. 82–97, 1936.

[20] G. K. Shortess and J. Krauskopf, "Role of involuntary eye movements in stereoscopic acuity," *JOSA*, vol. 51, no. 5, pp. 555–559, 1961.

[21] K. N. Ogle and M. P. Weil, "Stereoscopic vision and the duration of the stimulus," *Archives of Ophthalmology*, vol. 59, no. 1, pp. 4–17, 1958.

[22] G. Ashida and C. E. Carr, "Sound localization: Jeffress and beyond," *Current opinion in neurobiology*, vol. 21, no. 5, pp. 745–751, 2011.

[23] D. D. Stettler, A. Das, J. Bennett, and C. D. Gilbert, "Lateral connectivity and contextual interactions in macaque primary visual cortex," *Neuron*, vol. 36, no. 4, pp. 739–750, 2002.

[24] D. Y. Ts'o, C. D. Gilbert, and T. N. Wiesel, "Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis," *Journal of neuroscience*, vol. 6, no. 4, pp. 1160–1170, 1986.

[25] N. Sato and M. Yano, "A model of binocular stereopsis including a global consistency constraint," *Biological cybernetics*, vol. 82, no. 5, pp. 357–371, 2000.

[26] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[27] R. L. De Valois, E. W. Yund, and N. Hepler, "The orientation and direction selectivity of cells in macaque visual cortex," *Vision research*, vol. 22, no. 5, pp. 531–544, 1982.

[28] Y. Chen and N. Qian, "A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms," *Neural Computation*, vol. 16, no. 8, pp. 1545–1577, 2004.

[29] D. Green, "M. and swets, ja (1966). signal detection theory and psychophysics."

[30] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman, "Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors," *Science*, vol. 249, no. 4972, pp. 1037–1041, 1990.

[31] B. Julesz, "Foundations of cyclopean perception.," 1971.

[32] T. Trappenberg, *Fundamentals of computational neuroscience*. OUP Oxford, 2009.

[33] T. Taniai, Y. Matsushita, Y. Sato, and T. Naemura, "Continuous 3d label stereo matching using local expansion moves," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[34] Y. S. Heo, K. M. Lee, and S. U. Lee, "Robust stereo matching using adaptive normalized cross-correlation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 807–822, 2011.

[35] J. C. Murray, H. R. Erwin, and S. Wermter, "Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks," *Neural Networks*, vol. 22, no. 2, pp. 173–189, 2009.

[36] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. Chichilnisky, "The structure of multi-neuron firing patterns in primate retina," *Journal of Neuroscience*, vol. 26, no. 32, pp. 8254–8266, 2006.

[37] S. Grossberg and P. D. Howe, "A laminar cortical model of stereopsis and three-dimensional surface perception," *Vision research*, vol. 43, no. 7, pp. 801–829, 2003.

[38] W. Poelzleitner and D. P. Casasent, "Stereo disparity computation using gabor filters and feature selection techniques," in *Optical Pattern Recognition VII*, vol. 2752, pp. 268–281, International Society for Optics and Photonics, 1996.

[39] J. M. Harris and L. M. Wilcox, "The role of monocularly visible regions in depth and surface perception," *Vision research*, vol. 49, no. 22, pp. 2666–2685, 2009.

[40] J. J. Tsai and J. D. Victor, "Binocular depth perception from unpaired image points need not depend on scene organization," *Vision research*, vol. 45, no. 5, pp. 527–532, 2005.

[41] N. R. Goncalves, H. Ban, R. M. Sánchez-Panchuelo, S. T. Francis, D. Schluppeck, and A. E. Welchman, "7 tesla fmri reveals systematic functional organization for binocular disparity in dorsal visual cortex," *Journal of Neuroscience*, vol. 35, no. 7, pp. 3056–3072, 2015.

[42] D. B. Chklovskii, "Binocular disparity can explain the orientation of ocular dominance stripes in primate primary visual area (v1)," *Vision research*, vol. 40, no. 13, pp. 1765–1773, 2000.

[43] C. D. Gilbert and T. N. Wiesel, "Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex," *Journal of Neuroscience*, vol. 9, no. 7, pp. 2432–2442, 1989.

[44] M. Kinoshita and H. Komatsu, "Neural representation of the luminance and brightness of a uniform surface in the macaque primary visual cortex," *Journal of neurophysiology*, vol. 86, no. 5, pp. 2559–2570, 2001.

[45] C. Tyler and J. Foley, "Stereomovement suppression for transient disparity changes," *Perception*, vol. 3, no. 3, pp. 287–296, 1974.